APPROVAL SHEET

Title of Dissertation: Speech Input in Multimodal Environments:
Effects of Perceptual Structure on Speed, Accuracy,
and Acceptance

Name of Candidate: Michael A. Grasso
Doctor of Philosophy, 1997

Dissertation and Abstract Approved: _____
Dr. Timothy W. Finin
Professor
Computer Science and Electrical Engineering

Dissertation and Abstract Approved: _____
Dr. David S. Ebert
Assistant Professor
Computer Science and Electrical Engineering

Date Approved: _____

Curriculum Vitae

Name: Michael A. Grasso

Degree and Date to be Conferred: Ph.D., 1997

Professional Publications:

Michael A. Grasso, David Ebert, Tim Finin. The Effect of Perceptual Structure on Multimodal Speech Recognition Interfaces. ACM Transactions on Computer-Human Interaction, under review.

Michael A. Grasso, David Ebert, Tim Finin. Acceptance of a Speech Interface for Biomedical Data Collection. 1997 AMIA Annual Fall Symposium, under review.

Michael A. Grasso and Tim Finin. Task Integration in Multimodal Speech Recognition Environments. Crossroads, 3(3):19-22, Spring 1997.

Michael A. Grasso. Speech Input in Multimodal Environments: A Proposal to Study the Effects of Reference Visibility, Reference Number, and Task Integration. Technical Report TR CS-96-09, University of Maryland Baltimore County, Department of Computer Science and Electrical Engineering, 1996.

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

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

Professional Positions Held:

1988 - Present	President/Senior Computer Scientist. Segue Biomedical Computing, Laurel, Maryland.
1992 - Present	Instructor of Computer Science, Part-Time. University of Maryland Baltimore County, Department of Continuing Education.
1991 - 1993	Instructor of Computer Science, Part-Time. Howard Community College, Columbia, Maryland.
1987 - 1988	Senior Programmer/Analyst. Program Resources, Inc., Annapolis, Maryland.

1984 - 1985	Microbiology Technician and Computer Programmer. Johns Hopkins University, Baltimore, Maryland.
1981 - 1984	Medical Technologist and Microbiology Technician. Part-time positions and internships held while completing undergraduate education.

Abstract

Title of Dissertation: Speech Input in Multimodal Environments:
Effects of Perceptual Structure on Speed, Accuracy,
and Acceptance

Michael A. Grasso, Doctor of Philosophy, 1997

Dissertation Directed By: Dr. Timothy W. Finin, Professor
Computer Science and Electrical Engineering

Dr. David S. Ebert, Assistant Professor
Computer Science and Electrical Engineering

A framework of complementary behavior has been identified which maintains that direct manipulation and speech interface modalities 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. Based on this concept and the theory of perceptual structures, this work examined the hypothesis that the speed, accuracy, and acceptance of a multimodal speech and direct manipulation interface would increase when the modalities match the perceptual structure of the input attributes.

A software prototype to collect histopathology data was developed with two interfaces to test this hypothesis. The first interface used speech and direct manipulation in a way that did not match the perceptual structure of the attributes, while the second interface used speech and direct manipulation in a way that best matched the perceptual structure. A group of 20 clinical and veterinary pathologists evaluated the prototype in an

experimental setting using repeating measures. The independent variables were interface order and task order, and the dependent variables were task completion time, speech errors, mouse errors, diagnosis errors, and user acceptance.

The results of this experiment support the hypothesis that the perceptual structure of an input task is an important consideration when designing multimodal computer interfaces. Task completion time improved by 22.5%, speech errors were reduced by 36%, and user acceptance increased 6.7% with the computer interface that best matched the perceptual structure of the input attributes. Mouse errors increased slightly and diagnosis errors decreased slightly, but these were not statistically significant. There was no relationship between user acceptance and time, suggesting that speed is not the predominate factor in determining approval. User acceptance was related to speech recognition errors, suggesting that recognition accuracy is critical to user satisfaction. User acceptance was also shown to be related to domain errors, suggesting that the more domain expertise a person has, the more he or she will embrace the computer interface.

Speech Input in Multimodal Environments:

Effects of Perceptual Structure

on Speed, Accuracy, and Acceptance

by

Michael A. Grasso

Dissertation submitted to the faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements of the degree of
Doctor of Philosophy
1997

Dedication

To my parents, Silvio and Angela Grasso, who taught me to dream, and to my wife, Clare, for believing.

Acknowledgment

The completion of this dissertation was made possible through the support and cooperation of many individuals. Thanks to my advisors, Timothy W. Finin and David S. Ebert, who provided thoughtful guidance and encouragement through what seemed to be a never-ending process. Thanks also to Tulay Adali, Charles K. Nicholas, and Anthony F. Norcio, the other members of my committee, for helping to me to understand the significance of this research with respect to medical informatics, software engineering, and human-computer interaction, respectively.

Several people provided technical assistance. Judy Fetters from the National Center for Toxicological Research (NCTR) consulted with me on the software prototype and in the selection of tissue slides. Alan Warbritton from NCTR scanned the slides for the prototype. Lowell Groninger, Greg Trafton, and Clare Grasso helped with the experiment design and statistical analysis. The support staff at Speech Systems, Inc. helped with grammar development and answered countless technical questions.

Finally, thanks to 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. Special thanks to Dr. John Strandberg, Dr. Michael Lipsky, and Dr. Jules Berman for helping to identify participants.

Table of Contents

List of Tables

List of Figures

1. 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]. In addition, data entry has become the bottleneck of many applications in the field of medical informatics. This is due to hands-busy or eyes-busy restrictions during tasks such as patient care and microscopy.

An approach that addresses both of these limitations is to develop interface techniques 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 the use of speech in multimodal environments. To characterize the complementary behavior of speech and direct manipulation, several questions relating to the effects of reference visibility, reference predictability, reference number, and task integration are discussed. The specific focus of this effort is an empirical study of the effect of perceptual structure on the speed, accuracy, and acceptance of a multimodal speech and direct manipulation interface.

1.1 Speech Recognition Systems

Speech recognition systems provide computers with the ability to identify spoken words and phrases. Note that speech recognition focuses on word identification, not word understanding. The latter is part of natural language processing, which is a separate research area. This can be compared to entering characters into a computer using a keyboard. The computer can identify the characters which are typed. However, there is no implicit understanding by the computer as to what these characters mean.

1.1.1 Historical Perspective

The first speech recognition system was developed in 1952 on an analog computer using discrete speech to recognize the digits 0 through 9 with a speaker-dependent template matching algorithm [Davis, Biddulph, and Balashek 1953]. Recognition accuracy was reported to be 98%. Later that decade, a system with similar attributes was developed that recognized consonants and vowels [Dudley and Balashek 1958]. In the 1960s, research in speech recognition moved to digital computers, which became the basis for speech recognition technology to the present day [Lea 1993].

Despite rapid progress early on, limitations in computer architectures prevented any significant commercial speech recognition system development. Even though the data transfer rate of speech is only about 50 bits per second, the computational requirements associated with extracting this information are enormous. Over the last decade, a number of commercial systems have been successfully developed [Voice

Processing Magazine 1993]. However, since true natural language processing is still several years away, a successful speech-driven system must allow for restrictions in the current technology. These restrictions include speaker dependence, continuity of speech, and vocabulary size [Bergeron and Locke 1990; Peacocke and Graf 1990].

1.1.2 Speaker Dependence

Speaker-dependent systems are those which require some type of user training before they can be put to use. Speech recognition systems typically use a pattern matching algorithm, where the spoken words are compared with predefined templates to find the best match. Before this can occur, the user must create templates by saying each word in the vocabulary two or three times. Representative word phrases may also be read aloud, to identify how certain words will be spoken in context. A speech model consists of all the templates for a given vocabulary. Each operator of a speaker-dependent system must create a speech model by training the system to recognize his or her way of saying every word in the vocabulary. Depending on the vocabulary size, training can take from a few minutes to several hours.

Speaker-independent systems use generic models to recognize speech from any user. Generic models are created by combining existing templates from a variety of speakers. This approach is advantageous in that it does not require individual operators to train the system to recognize their voices. However, because the templates are not user-specific, accuracy rates are usually lower.

An alternative is the speaker-adaptive approach, which uses a generic model to eliminate initial training and then automatically generates user-specific models for each operator over time. Although initial training is eliminated, recognition accuracy is diminished until the system develops an adequate user-specific model.

1.1.3 Continuity of Speech

Continuous speech systems can recognize words spoken in a natural rhythm. Although this approach seems more desirable at first glance, continuous speech is harder to process because of the difficulty of identifying word boundaries - as in "youth in Asia" and "euthanasia." Variability in articulation, such as the tendency to drop consonants or blur distinctions between them - as in "want it" and “wanted" - can result in further misunderstanding. To increase accuracy, speech models for continuous speech systems include information on representative word combinations and context rules.

Isolated word systems require a deliberate pause between each word. Pausing after each word is unnatural and can be tiring. However, accuracy rates are usually higher with isolated word systems than with systems using continuous speech. Isolated systems are therefore thought to work best with vocabularies that consist mainly of individual command words.

1.1.4 Vocabulary Size

The vocabularies of various speech recognition systems can range from 20 to more than 40,000 words. Large vocabularies cause difficulties in maintaining accuracy,

but small vocabularies restrict the speaker. In addition, large vocabularies are likely to contain ambiguous words, which in speech recognition systems are words with pattern-matching templates that the computer will treat as similar - such as the words "tree" and "three."

Grammar rules can be added to impose constraints on the allowable sequences of words. These are especially important to offset technical limitations due to continuous speech or large vocabularies. A tightly constrained grammar is one in which only a small number of words can legally follow any given word, based on context of phrase structure. Keeping the list of candidate words small can increase recognition accuracy and decrease latency time during pattern matching, especially with large vocabularies. However, too many grammar rules can reduce the naturalness of communication.

1.1.5 Human Factors of Speech Interfaces

Along with technical characteristics of speech recognition systems, it is important to understand the human factors of speech as an interface modality. A criticism by Newell is that some researchers act as if the only bars to widespread adoption of speech interfaces are these technical limitations. Only occasional consideration is given to dialog design and other aspects necessary for an effective and efficient human interface [Newell 1992]. These comments highlight the importance of studying speech recognition interfaces as a human-computer interaction problem.

Speech is a unique modality with several profound qualitative differences from traditional user interface channels. The most significant is that speech is temporary. Once uttered, auditory information is no longer available to the user. This may place extra memory burdens on the user and severely limit the ability to scan, review and cross-reference information. A related limitation is that it is hard to represent spatial information, since the fleeting nature of speech makes it difficult to observe and manipulate the relative position of objects.

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, security and may add to environmental noise in the workplace.

Finally, more than other modalities, there is the possibility of anthropomorphism when using speech recognition. 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].

1.2 Direct Manipulation

Direct manipulation interfaces, made popular by the Apple Macintosh and Microsoft Windows graphical user interfaces, are based on a number of principles [Shneiderman 1993].

- Visual display of objects of interest.
- Selection by pointing, instead of typing.
- Rapid, incremental, and reversible actions.
- Immediate and continuous feedback of results and actions.

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. For example, the typical word processor today can display a document in its final form with fonts, graphics, and other characteristics exactly as they will appear when printed. Another example is the file manager that displays directories as a tree structure and files as icons.

In a direct manipulation environment, the computer is operated by direct engagement with the user interface. The commands themselves are physical actions, such as pointing, clicking, dragging, and sliding. For example, to delete a file, the user points to its icon and drags it to the trash can. Once the file is deleted, the user is given immediate confirmation by the fact that the file icon is no longer on the screen or that the trash can now appears to have something in it.

This approach has several potential advantages. The direct manipulation interface is based on intuitive metaphors with a consistent look-and-feel that enhances a user's ability to learn another program quickly. A hierarchy of menus makes available options

clear and minimizes the need to learn cryptic command languages. Users can immediately see the results of their actions, making error detection more natural and minimizing the need for error messages. Finally, users may gain more confidence and are more in control since they initiate commands by physical actions.

In contrast to this, arguments have been made that direct manipulation interfaces are inadequate for supporting transactions fundamental to applications such as word processing, CAD, and database queries [Buxton 1993; Cohen and Oviatt 1994]. 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. Shneiderman [1993] points to ambiguity in the meanings of icons and limitations in screen display space as problems with direct manipulation.

1.3 The Problem

It has been suggested that direct manipulation and speech recognition interfaces have complementary strengths and weaknesses which could be leveraged in multimodal user interfaces [Cohen and Oviatt 1994; House 1995; Cohen 1992]. By combining the two modalities, the strengths of one could be used to offset the weaknesses of the other. For simplicity, speech recognition will deal with the identification of spoken words, not necessarily natural language recognition, and direct manipulation will deal with mouse input.

The complementary advantages of direct manipulation and speech recognition are summarized in Table 1. Note that the advantages of one are the weaknesses of the other. For example, direct engagement provides an interactive environment which 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

Table 1: Complementary Strengths of Direct Manipulation and Speech

One of the key strengths of direct manipulation is that these physical commands are based on simple actions. One example of this are visual database interfaces based on the direct manipulation modality. An early example of a visual database interface is Query-by-Example [Zloof 1977], developed at IBM. Such interfaces rely on visual representations of the database structure, possibly with sliders and other mouse-driven interface objects to input query information [Ahlberg, Williamson, and Shneiderman 1992]. However, this method works best with databases consisting of well-formed ordinal data. Since the interface is directly tied to the actual underlying format of the

database, it is considered too low level [Cohen and Oviatt 1994]. In contrast to this, the declarative nature of speech recognition interfaces and their ability to use anaphoric references should make them more appropriate for complex actions.

The consistent look and feel of direct manipulation interfaces is believed to provide a foundation for allowing novices to learn the basic functionality of these programs quickly by generalizing the commonality between applications. The limitation of this approach is its increased dependence on the visual display of information. When there are only a few interface objects, it is easy to arrange them in a consistent manner. However, this approach quickly breaks down when there are dozens of interface objects to manipulate. Speech interfaces do not have such limitations, but the abstract characteristic of speech makes it difficult to employ the concept of look-and-feel in the same way.

Direct manipulation interfaces do not have problems with reference ambiguity. When the user selects an object, the computer will not misinterpret this selection as some other object. The down side to this is that there is only one way to reference an object. A problem with direct manipulation stated earlier is that not all objects have easily distinguishable references. In other words, while selecting an object is unambiguous to the computer, the actual meanings of these references may be obscure to the user. Speech interfaces have the opposite characteristic. Since objects can be referred to in multiple ways, the meanings of various references should be less ambiguous to the user. However,

due to recognition errors or grammar limitations, there is a greater chance the computer may not recognize this reference correctly.

Taking these observations into account, Cohen and Oviatt [1994] made the following statement with respect to the complementary benefits of direct manipulation and natural language. Note that this dissertation deals with the identification of words through speech recognition, not necessarily natural language interaction.

> *Theoretically, direct manipulation should be beneficial when the objects to be manipulated are on the screen, their identity is known, and there are not too many objects from which to select. Natural language interaction with computers offers potential benefits when users need to identify objects, actions, and events from sets too large to be displayed and/or examined individually and when users need to invoke actions at future times that must be described.*

For example, direct manipulation interfaces are 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 are better at specifying more complex actions when references are numerous and not visible. This is summarized in Table 2.

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

Table 2: Proposed Applications for Direct Manipulation and Speech

Based on these observations, a series of questions have been proposed to evaluate the effect of reference visibility, reference number, and task integration on the speed, accuracy and acceptance of direct manipulation and speech recognition systems. Such empirical results can be used to assist with the integration of speech with direction manipulation in multimodal environments. Due to time constraints, only the question on task integration was evaluated as part of this dissertation.

Relying on anecdotal arguments, expected results are that simple actions on a limited number of visible references would favor direct manipulation and complex actions on numerous, non-visible references would favor speech recognition. Intuitively, it is clear that direct manipulation interfaces are adversely affected by references which are not visible, since you must be able to see a reference in order to select it. In the same way, it is clear that speech recognition systems do not have this limitation, since any item can be referenced regardless of whether it is visible or not. Also, the declarative nature of speech recognition interfaces should allow the specification of more complex operations. However, this dissertation hypothesizes that this model of complementary behavior is only true under certain conditions related to the characteristics of the reference attributes and the type of interface task.

These original observations focused mainly on reference visibility. There are other attributes that may impact the speed, accuracy, and acceptance of both direct manipulation and speech recognition interfaces. The number of references is alluded to, however, only in the context of limiting visibility, such as when there are so many references that they all cannot be visible at the same time.

Also, regardless of reference attributes, the speed, accuracy, and acceptance may be impacted by how well the control structure of the input device matches the perceptual structure of the input task (whether the input attributes are perceived as integral or separable). It was reported that the performance of a unimodal, graphical interface improves when the structure of the perceptual space matches the control space of the input device [Jacob et al. 1994]. An appropriate follow-on question - and the focus of this study - is the effect of perceptual structure on multimodal tasks. A summary of reference attributes and interface tasks is in Table 3.

Reference Attributes	
Visible	The references are directly observable by the user and not obscured by other screen objects.
Numerous	The number of valid references available to the user are many.
Predictable	The references are sorted or otherwise familiar to the user.
Distinguishable	The references can be easily differentiated from each other.
Interface Tasks	
Integral	The input attributes cannot be attended to individually.
Simple	The task is implicit based on reference selection.
Spatial	The task is based on dimensional input.
Declarative	The task requires a description or anaphoric reference.
Computational	The task requires the input of numbers or formulas.

Table 3: Reference Attributes and Interface Tasks

1.4 Significance of this Study

There are three areas in which this research will contribute in a significant way to the understanding of speech recognition interfaces in human-computer interaction.

1. Replace anecdotal arguments with scientific evidence.
2. Identify situations where speech is the preferred modality.
3. Increase our understanding of speech in multimodal environments.
4. Address the data entry bottleneck in medical informatics.

The literature is filled with anecdotal arguments about the applicability of speech recognition interface. Shneiderman [1992] points out four such areas: when the hands are busy, the eyes are busy, mobility is required, and in harsh environments. Cohen and

Oviatt [1994] suggest a similar set of conditions: when the user's hands or eyes are busy, only a limited keyboard or screen is available, the user is disabled, pronunciation is the subject matter of computer use, and when natural language interaction is preferred.

The first area where this research will contribute is to help replace these anecdotal arguments on the applicability of speech and the complementary advantages of direct manipulation and speech recognition with scientific evidence. Such a framework for research in human-computer interaction has been identified by Shneiderman [1993] as a foundational approach. By emphasizing controlled experiments which yield more objective and reliable results, arguments about "user friendly systems" are replaced with a more scientific approach.

The second area where this research will contribute is by identifying those situations where speech is the preferred interface modality. Note that the anecdotal arguments on the applicability of speech, while intuitive, have a particular bias. That is, they imply that speech is always a second choice that is only appropriate when traditional keyboard and screen interfaces are impractical. While acknowledging this bias, 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. Cohen and Oviatt [1994] state that no principled methods exist which can predict those circumstances where speech will be the most effective, efficient, or the preferred interface modality. Still others point out that there is still a lack of theoretical work and empirical results [Carbonell 1994], and the need for

rigorous scientific investigation into the applicability of speech as an interface medium [Damper 1993].

The third area where this research will contribute is by increasing our understanding as to when and under what conditions speech can be integrated with mouse input in multimodal environments. Cole et al. [1995] note the role that spoken language should ultimately play in multimodal systems is not well understood and calls 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. The focus of this research is user perception of the input task based on the theory of perceptual structures. Such research is needed to understand how people select and integrate different modalities in the context of different types of human-computer interaction [Oviatt and Olsen 1994].

The objective of this dissertation was to the study the effect of the perceptual structure of multidimensional input tasks on the speed, accuracy and acceptance of multimodal direct manipulation and speech recognition systems. Such empirical results can be used to assist with the integration of speech in multimodal environments.

The fourth area where this research will contribute is by addressing the data entry bottleneck in medical informatics [Grasso and Grasso 1994; Dillon, McDowell, Norcio, DeHaemer 1994; McMillan and Harris 1990]. Histopathologic data collection in animal toxicology studies was chosen as the application domain for user testing. It includes several significant hands-busy and eyes-busy restrictions. It is based on a highly

structured, specialized, and moderately sized vocabulary based on 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. Also, the input tasks mainly involve reference identification, with little declarative, spatial, or computational data entry required, which should eliminate any built-in bias toward either modality.

1.5 Research Questions

The three proposed studies are based on the following three research questions. Included with each research question is a summary of the literature review from Section 1, predicted results, and null hypotheses for statistical evaluation.

Only question number one on task integration has been studied as part of this doctoral research project. The other two questions were discussed, but not studied. This was to ensure that this research effort was completed in a reasonable amount of time.

Question 1

What multidimensional tasks can best be integrated with multimodal speech and direct manipulation?

Literature

The performance of multidimensional, unimodal input tasks is affected by whether the dimensions are perceived as integral or separable. Users are more likely to switch from one modality to another when there is a change in functionality or context.

Predicted Results

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.

Null Hypothesis 1

The integrality of input attributes has no effect on the speed of the user.

Null Hypothesis 2

The integrality of input attributes has no effect on the accuracy of the user.

Null Hypothesis 3

The integrality of input attributes has no effect on acceptance by the user.

Question 2

How does the lack of visible references affect the speed, accuracy, and acceptance of speech and direct manipulation interfaces?

Literature

Direct manipulation interfaces perform better with visible references while speech interfaces perform better with non-visible references.

Predicted Results

Decreasing visibility has a negative impact on the speed, accuracy, and acceptance of both direct manipulation and speech interfaces. The negative impact on speech interfaces is greater than or equal to that of direct manipulation, except when those references have a high degree of predictability.

Null Hypothesis 4

Reference visibility has no effect on the speed of the user.

Null Hypothesis 5

Reference visibility has no effect on the accuracy of the user.

Null Hypothesis 6

Reference visibility has no effect on acceptance by the user.

Question 3

How does increasing the number of references affect the speed, accuracy, and acceptance of speech and direct manipulation interfaces?

Literature

Direct manipulation interfaces perform better with fewer references while speech interfaces perform better when there are numerous references.

Predicted Results

Increasing the number of references has a negative impact on the speed, accuracy, and acceptance of both direct manipulation and speech interfaces. The negative impact on speech interfaces is greater than or equal to that of direct manipulation, except when those references have a high degree of predictability.

Null Hypothesis 7

The number of references has no effect on the speed of the user.

Null Hypothesis 8

The number of references has no effect on the accuracy of the user.

Null Hypothesis 9

The number of references has no effect on acceptance by the user.

2. Literature Survey

This chapter contains a review of literature regarding this research effort. Related work in multimodal interfaces using direct manipulation and speech recognition is covered. An overview of research concerning key reference attributes and interface tasks is included. Motivations for the application of speech interfaces in the biomedical area are presented. Background information on the target application of data collection in animal toxicology studies is given. The chapter concludes with an outline of preliminary work in biomedical speech interfaces.

2.1 Multimodal Speech Recognition Interfaces

Several research efforts have attempted to develop multimodal interfaces using direct manipulation and speech recognition. Three of these are described below. While each has a different area of emphasis, all three are feasibility studies centered around the development and testing of a multimodal interface to demonstrate proof-of-concept. In their conclusions, they all call for empirical evaluations to refine and evaluate these interface techniques.

Two approaches to multimodal interfaces are presented - synergistic and integrated. Both are shown graphically in Figure 1. In a synergistic interface, each modality can perform the same set of tasks. No new functionality is added to the interface, except that the user can select the input device which is most convenient at any

given time. One example of this is navigation under Microsoft Windows, where either the mouse or the keyboard can be used to switch to the other windows. In contrast to this, an integrated interface is different in that there are certain tasks which can only be carried out by using both input devices together. The advantage here is that the functionality of the interface is extended with integrated tasks like "point-and-speak."

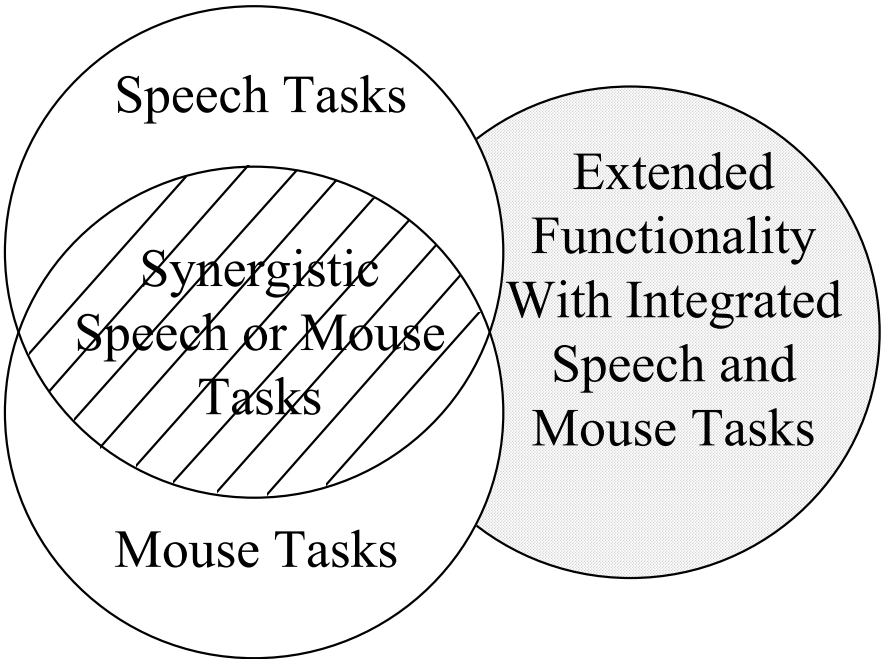

Figure 1: Synergistic versus Integrated Interface Tasks

2.1.1 Multimodal Access to the World-Wide Web

One effort at the Oregon Graduate Institute sought to evaluate spoken language as an alternative interface to multimedia applications [House 1995]. Specifically, a

multimodal interface to the World-Wide Web [CERN] was developed. The basic architecture of the system was a remote recognition-capable Web server with speech recognition software and speech-capable HTML (Hypertext Markup Language) documents. The local Web browser was extended to digitize the user's utterances and send them to the server for speech recognition and processing.

It was noted that, while the mouse-based interface can be credited with much of the popularity of the Web, there are inherent limitations. These limitations focus on difficulty in performing complex commands and access to documents that cannot be reached by a visible link. The latter - access to non-visible references - was the focus of their effort, and was motivated by the framework for complementary behavior between natural language and direct manipulation suggested by Cohen [1992].

Speech and direct manipulation were both used to develop an interface with a synergistic interaction style that allowed either modality to perform the same set of tasks [Lefebvre, Duncan, and Poirier 1993]. The user could then select the input device that best suits the task at hand. Speech recognition and direct manipulation were used as complementary modalities. As a result, speech input was believed to allow access to information that was not directly available with mouse-based systems, such as navigating to HTML links that were not visible.

The main advantage of using speech is that all references are potentially available, even when they are not visible. One question not raised was the need for predictability when selecting non-visible references. For example, will users have a

difficult time selecting a non-visible reference if they do not know, or cannot predict, what those references are? This may be especially true with the World-Wide Web, as users navigate through unfamiliar documents searching for information.

Another difficulty is the use of multiple labels like “Click here,” which can result in reference ambiguity. In addition, the use of Postscript, and other presentation-based encodings which assume a single display format, limit the ability to use speech output on the Web [Ramon 1995]. These factors highlight the need to enforce document development guidelines [Conte 1994] before speech-driven Web access can become commonplace.

2.1.2 Integrated Multimodal Interface

An alternative approach is to use natural language and direct manipulation to develop what has been called an integrated user interface. In one such effort, Cohen [1992] attempted to not simply provide two or more separate modalities with the same functionality, but to integrate them to produce a more productive interface. For example, along with traditional unimodal operations like “point-and-click,” there can be integrated ones like “point-and-speak.” The guiding principle in this research was to use the strengths of one modality to overcome the weaknesses of the other.

For simplicity, the term “natural language” was used independent of the transmission medium - keyboard, speech, or handwriting. Even though meaningful differences exist between spoken, keyboard, and written interaction [Oviatt and Cohen

1991; Kassel 1995], those differences were not germane to the key point about modality integration. Note also that this dissertation deals with speech recognition, not natural language, both of which have distinct characteristics [Shneiderman 1980]. However, there is enough overlap between the two for this research to provide relevant background material.

Based on this objective, a prototype multimodal system was developed using an integrated direct manipulation and natural language interface. Several examples were cited where the combination of language and mouse input together were thought to be more productive than either modality alone. For example, natural language allowed the use of anaphoric references (pronouns). However, the exact meaning of these references can be ambiguous. Following Webber's arguments [1986], the prototype used icons to explicitly display what it believed the valid references were, given the current context. The combination of anaphoric reference with pointing used the unambiguous nature of pointing to overcome this error-prone aspect of natural language processing.

A second example of integration introduced by Cohen was with the use of time. One might assume that direct manipulation would be better than speech for dealing with time by using a slider bar as a graphical rendition of a time line. However, this is not always the case. Finding timed events with a slider can be an extremely slow linear search process, especially if there is a large range of time intervals to scan. If the granularity of the slider is too large, selecting the exact time event may not be possible. Also, sliders typically allow the selection of only one time point. To overcome these

limitations, the prototype used natural language to describe the times of interest. The prototype then composed a menu of all time points selected with the slider set to the first one found. Here, natural language was used to overcome a weakness in direct manipulation - the selection of unknown objects (in this case, time points) from a large set.

Using the mouse to disambiguate the context of speech input has also been explored by the Boeing Company [Salisbury et al. 1990]. Their motivation was that human communication is multidimensional and that conversations include more than just spoken words. The combination of graphics and verbal data to complete or disambiguate the other was termed “talk and draw.” Within this framework, operators would input requests by speaking commands while simultaneously selecting graphical objects with a mouse to determine the context of these commands.

2.1.3 Multimodal Window Navigation

A project at the Massachusetts Institute of Technology used speech as an auxiliary channel to support window navigation [Schmandt, Ackerman, and Hindus 1990]. Xspeak provided a speech interface to X Windows by allowing navigational tasks usually performed with a mouse to be controlled by speech instead. The effort was developed with the assumption that speech input is more valuable when it is combined with other input devices and that most successful speech recognition systems have small vocabularies, are speaker-dependent, and use discrete speech.

The X Windows system uses a spatial metaphor to organize applications on a monitor in three dimensions. However, it uses a two-dimensional device for window navigation, namely the mouse. When there are many overlapping windows, it can be difficult to reach some applications directly with the mouse. Xspeak was therefore designed to improve navigation in this type of environment. Each window is associated with a voice template. When the word represented by a template is spoken, the window is moved to the foreground and the mouse pointer is moved to the middle of the window. Window navigation can be viewed as a hands-busy task. Using Xspeak, users can manage a number of windows without removing their hands from the keyboard.

Initial testing revealed that while speech was not faster than the mouse for simple change-of-focus tasks, the advantage shifted toward speech if the desired window was partly or completely hidden. Another observation was that the users most inclined to choose speech input increased the number of overlapping windows or the degree of overlap.

2.2 Reference Attributes

The following section discusses how reference visibility, reference number, and reference predictability can affect the performance of speech interfaces.

2.2.1 Reference Visibility

It has been suggested that speech input is better than mouse input when selecting non-visible references. However, due to the fleeting nature of spoken words, the impact of non-visible references on the cognitive costs of the user must be considered. For example, the less feedback or prompting a program provides, the more a user has to remember, and the more performance may suffer. The following studies suggest that the lack of visible references has a negative cognitive impact on both speech and direct manipulation interfaces.

An experiment was conducted at the University of Maryland College Park to demonstrate the utility of speech input for command activation during word processing [Karl, Pettey, and Shneiderman 1992]. It was believed that speech would be superior to the mouse with respect to the activation of commands. Also, word processing was considered a hands-busy, eyes-busy application, since the user would have to interrupt typing of text in order to execute word processing commands. Speech-activated commands were found to be faster than mouse-activated commands and to have similar error rates. Speech showed the greatest advantages during command-intensive tasks as opposed to typing-intensive tasks.

One unexpected result was that subjects made significantly more memorization errors when using speech. For one of the tasks, not all of the information could be displayed on the screen at one time. This meant that the participants had to memorize symbols and page up and down while using speech-activated commands. The researchers

observed a less-than-expected performance increase for this task using speech. When questioning the users, at least half noted that it was harder to memorize and recall descriptions when using voice input. Memorization problems did not interfere with mouse users performing the same task. This finding might explain why the use of graphics to display the visual context in which the various grammatical rules applied was shown to improve the speed and accuracy of speech recognition [Wulfman et al. 1993].

Another study observed increased cognitive requirements while retrieving hidden information with a mouse [Wright, Lickorish, and Milroy 1994]. To conserve space, a common practice is to remove information from computer displays that readers will only need intermittently. This information is often accessible by a single mouse click. The study demonstrated that this practice impairs one's memory for other task components due to increased cognitive costs. These findings suggest that software should be designed with additional memory support for users with small screens and also help to explain the success of icon bars and ribbon displays which give people immediate access to the functions they frequently use.

A related study empirically evaluated the effect of various user-interface characteristics on data entry performance for clinical data [Poon and Fagan 1994]. The characteristics tested were 1) displaying results as one long scrolling list or as a series of pages, 2) using dynamic palettes which pop-up when needed and are customized to the particular data collection event or fixed palettes, and 3) showing all findings or just those which are relevant in the current context.

Intuitively, one can argue that the use of scrolling, dynamic palettes, and showing relevant results allows for greater flexibility and better management of screen space. However, all three had a negative impact on performance due to increased memory requirements on the user. The study found that paging, fixed palettes, and showing all results provided better performance. With these characteristics, users could memorize the screen position of various objects and the need for commands to explicitly invoke or dismiss dynamic palettes was eliminated. Also, by showing all results, not just relevant ones, users were more confident of their findings and spent less time with follow-up questions.

The use of scrolling, dynamic palettes, and relevant findings resulted in a user interface with more variation than their counterparts. This, in turn, increased the cognitive costs on the user and decreased performance. A similar conclusion was reached by Mitchell and Shneiderman [1989]. This effort set out to show that dynamic or adaptive menus would perform better than fixed menus. Instead, they discovered that frequent changes to the menu order have a negative effect on users. They concluded that stability and predictability in menus was the preferred approach.

2.2.2 Vocabulary Size

It has been suggested that the more references there are (or the larger the vocabulary), the better suited an application may be to speech recognition [Cohen and Oviatt 1994]. While this might be the case, there seems to be little evidence to support

this. Consider the task of selecting an item out of a list, or a 1-out-of-N task. For small lists, Welch [1977] showed that the entry of numbers using a keyboard is faster and less error prone than entry by speech. This was later confirmed by Damper [1993].

In contrast to this, increasing vocabulary size had an interesting effect in the synergistic multimodal window navigation project described earlier [Schmandt, Ackerman, and Hindus 1990]. When there were fewer windows, the mouse performed better than speech. However, the more windows there were (or the larger the vocabulary), the more speech outperformed direct manipulation. There is another reference attribute to consider other than the size of the vocabulary or the lack of visibility. Note that increasing the number of references did not adversely affect the performance of the speech interface. However, since each window was given a name by the user, there should also have been a high degree of predictability within the vocabulary.

Dillon evaluated the effect of vocabulary size among nurses during a hands-busy data entry task. He showed that a larger inclusive vocabulary can lead to far fewer non-recognized phrases [Dillon, Norcio, DeHaemer 1993]. Although one vocabulary was larger than the other, both were functionally equivalent. The larger vocabulary contained alternative word choices while the smaller one used a minimal set. With both vocabularies, the user still had the same number of functional tasks to consider. This suggests that broadening a vocabulary to accommodate alternative phrases should increase the performance of a speech interface. However, it does not imply that increasing vocabulary size by adding functionality of a vocabulary will do the same.

2.3 Multimodal Input Tasks

An area of growing interest is to identify the best ways to integrate speech into multimodal environments. Research here includes those conditions where people are likely to integrate two input modalities as well as what advantages can be leveraged. Two such efforts are presented. The first studied how the perceptual structure of the input attributes can affect the performance of multidimensional input tasks. An overview on the perception of structure is given as background material. The second examines those conditions under which a person is likely to combine two modalities.

In this section, it is important to understand the difference between "integral" and "integrated," since they sound similar but have different connotations. The term, "integral," is used in the theory of perceptual structure to characterize the relationship between the dimensions of a structure as indivisible. This can refer to the structure of an input device or an input task. The term, "integrate" is used to describe the combining of two modalities and using them in concert.

2.3.1 Theory of Perceptual Structures

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 peculiar 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.

Perception occurs in the head, somewhere between the observable stimulus and response. Perception consists of various kinds of processing that have distinct costs, so the response is not just a simple representation of the 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.

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.

The next two examples are from Pomerantz and Lockhead [1991]. Consider two sequences: XOXOXOXOXO and OXXXOOXOXO. Each has five Xs and five Os and each is equally likely to occur from the 1,024 possible patterns. Yet the first pattern is considered better than the second, because of inferred subsets. The first pattern has fewer inferred alternatives than the second because it is perceived as more regular and predictable than the second. The goodness of a pattern is correlated with redundancy. Good stimuli are perceived as being in small subsets. The more redundancy, the smaller the subset. Given two subsets, each created from different total sets of the same size, if one subset is smaller, it has more redundancy. Also, by observing a single stimuli, we may be able to infer what the subset is. For example, given the letter E, one may infer the subset included letters of the alphabet.

There are two ways to introduce structure into a system. One is to present the stimuli in a nonrandom order, such as repeating a sequence of five circles in the same order. The other is to correlate the dimensions of a structure, such as an increase in circle size corresponding to an increase in its color or lightness.

The introduction of structure can improve performance, as shown by the following example. Consider a set of five circles that vary in size, and a set of twenty-five circles that vary in size and lightness. The one-dimensional circles are a 1 x 5 set while the two-dimensional circles are a 5 x 5 set. The 1 x 5 set should have performance advantages, due to its smaller size. However, by adding structure, this advantage is eliminated. Structure can be added by correlating the two attributes of the 5 x 5 set. In

this arrangement, an increase in size corresponds to an increase in lightness for each of the five sizes. The result is that the 5 x 5 set would now have only five valid choices, just like the 1 x 5 set.

2.3.2 Integrality of Input Devices

Speech and the mouse as input devices have significantly different control structures. The following study suggests that this can have a measurable impact on performance based on whether the control structure of each device matches the perceptual structure of the input task. Therefore any consideration of the advantages of one modality over the other should take into account these differences.

In this study, the researchers 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 input tasks with three inputs each were used, 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 is 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 (or how these dimensions are perceived) of an input task is key to the performance of multidimensional input devices on multidimensional tasks.

Consider the graphical input tasks in Table 4. Both use three attributes. However, Garner [1974] has shown that the attributes of the first graphical task are integral. That is, all three dimensions are in the same perceptual space. With the other graphical task, the three attributes are in separate perceptual spaces. This effort focused on multidimensional input on unimodal input devices. For multimodal environments, an appropriate follow-on question is the effect of integral and separable tasks using two or more input modalities in concert. Along with the graphical tasks, Table 4 contains integral and separable tasks from the biomedical application domain used in this dissertation.

Domain	***Task Type***	***Input Attributes***
Graphical	Integral	Location and size of a screen object
	Separable	Location and color of a screen object
Biomedical	Integral	Qualifier and morphology (marked inflammation)
	Separable	Site and qualifier (follicle marked)

Table 4: Integral and Separable Input Attributes

2.3.3 Integrating Input Modalities

A number of related studies were performed to examine how people might integrate input from different devices in a multimodal computer interface. The first study

used a simulated service transaction system with verbal, temporal, and computational input tasks using both structured and unstructured interactions [Oviatt and Olsen 1994]. Participants were free to use either handwriting, speech, or both during testing. The following results were reported. As shown in Table 5, 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.

Task	***Written***	***Spoken***
Verbal/Temporal	13.0%	87.0%
Verbal/Temporal & Computational	18.0%	82.0%
Textual	9.7%	90.3%
Textual & Computational	14.7%	85.3%
Proper Names	21.5%	78.5%
Structured	6.9%	93.1%
Unstructured	18.9%	81.1%

Table 5: Ratio of Written to Total Input

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 patterns identified in Table 6. Also shown in Table 6 is the tendency toward

certain combinations, such as written data and spoken command versus spoken data with written command.

Task	***Occurrence***
Written Input and Spoken Correction	50%
Spoken Input and Written Correction	50%
Written Data and Spoken Command	73%
Spoken Data and Written Command	27%
Spoken Text and Written Digits	85%
Written Text and Spoken Digits	15%

Table 6: Contrastive Pattern of Modality Use

A related study examined the use of spoken and written input while interacting with an interactive map system [Oviatt 1996]. Input modality (speech, writing, multimodal) and map display format (structured, unstructured) were manipulated in a simulated environment to measure performance errors, spontaneous disfluencies, and task completion time. With the previous study predicting users would prefer multimodal to unimodal interfaces, this study explored whether there were performance advantages as well. A simulated service transaction system was used by participants to assist with map-based tasks.

The study revealed that increased length of spoken utterances and unstructured displays resulted in more 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 that used pen-based pointing and gestures to identify location and speech for other data input.

The two key points in this section on multimodal input tasks are the positive relationship of contrastive functionality to multimodal interaction and the application of the theory of perceptual structures to multidimensional, unimodal input tasks. These findings were used to develop the dissertation hypothesis that multidimensional, multimodal input tasks will exhibit increased speed, accuracy, and acceptance when the input attributes are perceived separable. When the attributes are integral, unimodal input would be more beneficial.

2.4 Motivations of Speech in Medical Informatics

Automated speech recognition can address two key concerns in human-computer interaction: the demand for ease of use and constraints on the user's ability to work with the keyboard or mouse. The technology is still limited, however, with most successful systems using small to medium-size vocabularies with well-defined grammar rules. In the area of medical informatics, the main applications of speech recognition systems described in the literature are for 1) template-based reporting, 2) natural language processing, 3) multimodal integration of speech with other methods of input, and 4) data entry in hands-busy environments. The first two reflect the need for more intuitive

interfaces. The latter two deal with limitations of traditional input using the keyboard or mouse.

2.4.1 Template-Based Reporting

Template-based reporting systems have been used in radiology, pathology, endoscopy, and emergency medicine. They have large vocabularies, recognize discrete speech, and are speaker-adaptive systems designed to generate template-based reports using fill-in forms, trigger phrases, and free-form speech. Turnaround time is decreased and accuracy is increased by eliminating the need for dictation and transcription by clerical personnel.

Reaction to this approach has been mixed. For autopsy pathology, it was noted that a greater degree of computer literacy is required and that the need for typed input is not eliminated [Klatt 1991]. When applied to endoscopy, the process took longer than standard dictation and nevertheless collected less information [Massey, Geenen, and Hogan 1991]. These problems were attributed to the fact that therapeutic endoscopic procedures are complex and not suited to a template-based reporting format. The free-form speech method, in which single words are printed as they are spoken, was found to be too slow to be useful [Dershaw 1988]. This was probably due to increased computational requirements associated with larger vocabularies (up to 40,000 words). On the positive side, the formality of the process seemed to provide other benefits. One researcher noted that 80% of emergency room reports were adequately completed with a

speech recognition system, as compared with 30% when reports were dictated or handwritten records were used [Hollbrook 1992].

2.4.2 Natural Language Processing

A group at Stanford University studied the use of speech as an improved interface for medical systems. Initial work focused on the development of three prototype speech-driven interfaces [Issacs et al. 1993] along with research on how clinicians would like to speak to a medical decision-support system [Wulfman et al. 1993]. It was noted that the use of template-based dictation with fill-in forms worked well only when the documentation task was limited to a few standardized reports. Template-based reporting may be inadequate in clinical domains, because the required documentation is less standardized. At the same time, current speech recognition technology does not permit the processing of free-form natural language. Methods that circumvent shortcomings in the current technology while maintaining the flexibility and naturalness of speech are being explored.

Three prototype systems were developed that were more complex linguistically than template-based reporting, and the typical entries could not easily be selected from a simple presentation of menus. The systems had a speaker-independent vocabulary of more than 38,000 words using continuous speech. In addition, Windows-based graphics were used as control and feedback mechanisms for the various grammatical rules in the system. This use of graphics to display the visual context in which the various

grammatical rules applied was shown to improve the speed and accuracy of recognition except when the grammar was complex. Overall, the evidence suggests that graphical guidance can be used effectively when the vocabulary is sufficiently constrained.

2.4.3 Speech in Multimodal Environments

A different approach for speech recognition is to develop multimodal systems that use speech in combination with other input devices. The goal in this case is not to replace the keyboard or mouse but to simplify or accelerate the input process. One such system, designed to assist in the collection of stereological data, combined speech input with a digitizing pad [McMillan and Harris 1990]. Each data set consisted of an object name recorded by voice, followed by X and Y coordinates entered with a digitizing pad. The system was used for boundary analysis and histomorphometry of bone and skin. It had a small speaker-dependent vocabulary (less than 50 words) for object names and voice commands, and recognized discrete speech. The system allowed a user to choose between a small set of control words and about 20 object names. The combination of speech and a digitizing pad was shown to accelerate the data collection process.

2.4.4 Hands-Busy Data Collection

Several efforts used a speech-driven approach to facilitate the collection of data in a hands-busy environment. This has been a key motivation for the application of speech in the medical area as well as in other domains. Hands and eyes-busy data collection was also the principal motivation behind the preliminary work described below.

One study examined the feasibility of using speech recognition to record clinical data during dental examinations [Feldman and Stevens 1990]. Systems of this type would eliminate the need for a dental assistant to record results. Speech input was shown to be slower. However, when the time needed to transfer results recorded by the dental assistant into the computer was considered, the speech method was considered faster. Speech input also had more errors, although the difference was not statistically significant. Overall, the study suggested that speech recognition may be a viable alternative to traditional charting methods.

Another effort designed a speech interface for an anesthetist's record keeping system [Smith et al. 1990]. Anesthetists are responsible for recording information on drugs administered during medical procedures. Due to hands-busy limitations, a long interval typically exists between an event and its recording, which can compromise the completeness and accuracy of the manual record. By using speech input, this data can be collected during the medical procedure, while the anesthetist's hands are busy. The system used a vocabulary of around 300 words. Preliminary testing showed an accuracy rate of 96%, even in a noisy operating room.

Hands-busy data collection has also been applied to the analysis of bone scintigraphic data [Ikerira et al. 1990]. Such diagrams are analyzed to study metastases of malignant tumors. A speech system was developed to allow doctors to enter the results of image readings into the computer while looking at the images instead of the terminal. In 580 voice-entered reports, response time was shortened in comparison with dictation or writing by hand.

2.5 Data Collection in Animal Toxicology Studies

Data entry has become the bottleneck of many scientific applications designed to collect and manage information related to experimental studies. In animal toxicology studies, this is true because of the need to collect data in hands-busy or eyes-busy environments. For example, during microscopy, the operator's hands and eyes are occupied with the process of examining tissue slides. During necropsy, gross observations and organ weights must be collected while the operator's hands are busy and soiled. With in-life data collection, technicians record daily observations while handling animals. An ancillary data collection issue is that it may not be practical to keep computer equipment in animal rooms and laboratories, where it is most convenient to record observations.

Large volumes of pathology data are processed during animal toxicology studies. These studies are used to evaluate the long-term, low-dose effects of potentially toxic substances, including carcinogens. This information must be collected, managed, and

analyzed according to Good Laboratory Practice regulations for animal studies [U.S. FDA 1978]. Since the 1970's, several systems have been developed to automate this process [Cranmer et al. 1978; Faccini and Naylor 1979]. Procedures for manual data entry were set up. Others included interfaces to clinical chemistry and hematology analyzers to automate data collection [Daly et al. 1989]. Today, however, the collection of microscopic, gross, and in-life observations is still a limiting factor, due to hands-busy and eyes-busy restrictions.

Several software systems have been developed in this area, such as the Toxicology Data Management System (NCTR, Jefferson, AK), Starpath (Graham Labs, San Antonio, TX), and Labcat (Innovative Programming Associates, Princeton, NJ). These and other applications deal with specific information management and analysis issues. However, automation at the source of data collection through speech recognition has yet to be fully explored. Speech is a natural means of communication that would address the data entry bottlenecks which can occur with standard data collection processes. The highly structured and moderately sized vocabulary (as opposed to free-form and large vocabulary) required by these applications can easily be supported by current speech recognition systems. Automating at the source of data collection has the potential to greatly reduce transcription and data validation costs that consume 25 to 33 percent of the total cost of bringing new drugs to market [Green 1993].

2.6 Preliminary Work

Preliminary work by the author in this area includes a feasibility study of voice-driven data collection [Grasso and Grasso 1994]. The objective was to determine the feasibility of using voice recognition technology to enable hands-free and eyes-free collection of data related to animal toxicology studies. A prototype system was developed to facilitate the collection of histopathology data using only speech input and computer-generated speech responses. After testing the prototype system, the results were evaluated to determine the feasibility of this approach and provide a basis for implementing voice-driven systems that support microscopic, gross, and in-life data collection.

2.6.1 Materials

The hardware for this study consisted of an IBM-compatible 486/33 computer with Microsoft Windows 3.1 (Redmond, WA). Software was developed under Microsoft Windows using Borland C++ 3.1 and the Borland Object Windows Library 1.0 (Borland International, Inc., Scotts Valley, CA). Watcom SQL for Windows 3.1 (Watcom International Corporation, Waterloo, Ontario, Canada) was the relational database chosen. The Verbex 6000 AT31 Model 0637 Voice Input Module with 3 megabytes of memory, 40 MHz processor, and text-to-speech synthesis was used for voice recognition and computer-generated voice responses (Verbex Voice Systems, Inc., Edison, NJ).

Two separate interfaces were developed for data collection. One used the keyboard, mouse and computer monitor with standard interface objects such as dialog boxes, push buttons, and pulldown menus. The other used only speech input and computer-generated speech responses with no visual feedback. Note that these were two distinct user interfaces and that speech-driven capabilities were not merely added to the Windows user interface. Simply adding speech to an existing user interface has been shown to decrease system integrity or cause integration discontinuity [Wulfman et al. 1988].

The grammar was a continuous-speech, speaker-dependent vocabulary of 900 words, based on the Pathology Code Table [1985]. The list of possible words and phrases was divided into functional subsets for navigation, voice response, error correction, nomenclature terms, and data collection.

2.6.2 Methods

An informal series of four tests was conducted. In each test, the subject was either a pathology assistant, medical technologist or software engineer. The first test was to train the system to recognize each user's voice by reading each word twice, followed by reading representative words in context.

The second test was used to validate the accuracy of the voice recognition system apart from the application. Each user was asked to read a series of 100 randomly generated phrases. The number of correctly recognized phrases was used to compute the

recognition accuracy. If a phrase was accidentally read incorrectly, it was not counted as an error, and the user was given a second chance to read the phrase again. Invariably, users needed to repeat the training for specific words that the system was not recognizing consistently. If retraining was successful, these were not counted as errors.

In the next test, each user was asked to navigate to various animals and enter several microscopic observations. Here, voice recognition was not used. Instead, the keyboard and mouse were used for input and a computer monitor for visual responses. This test was to allow the users to familiarize themselves with the environment and provide a comparison for data entry using voice input.

The final test required each user again to navigate to various animals and enter several microscopic observations. This time, however, the mouse, keyboard, and monitor could not be used. Instead, each user relied on voice input and computer-generated voice output.

2.6.3 Results and Discussion

Overhead associated with training was a limiting factor. Roughly four to eight hours were required for each user to train on the entire vocabulary of 900 terms. The mean recognition rate was 97% in the accuracy test. In the last test, most participants felt uncomfortable at first when entering observations without any visual feedback. This was due in part to difficulty in understanding computer-generated speech. After a few practice runs, they were entering data without assistance. However, many felt the system should

provide more feedback during data collection - be it visual or audible. The mean recognition rate in this test was also 97%. These accuracy rates were determined under controlled conditions, so they should be viewed as a best-case scenario.

The initial training requirements are a potential hindrance to the acceptance of a system of this type. In a time when few people, if any, read the user's guide, it is difficult to envision a pathologist spending hours training the system to recognize his or her voice. An alternative that warrants further study is a speaker adaptive approach. Here, instead of training the system, operators would use a set of generic voice recognition templates, which would automatically be adapted for each person with continued use.

Another interesting observation has to do with word conflicts in the vocabulary. Such conflicts can occur with short, similar sounding words like "tree" and "three". It was initially believed that a vocabulary of complex medical terms would be immune to such problems. However, there were some conflicts with phrases like "inferior vena cava" and "superior vena cava".

The area of computer feedback requires additional research. Since the system operated in an eyes-busy environment, there could be no visual computer feedback. Several areas were anticipated where audible confirmation would be appropriate, such as when a word was recognized by the system or when an observation was saved in the database. However, occasionally there were moments of "dead air time" when the computer was involved in a large database transaction or the speech recognizer was parsing a complex phrase. Here, it might have helped to provide additional audible

feedback so the user knew when the computer was busy, similar to displaying an hourglass cursor on the computer monitor when a program is busy. This is not always easy to do. For instance, a software application can only determine when a recognizer event ended, not when it began, which makes it difficult to know when to transmit a busy signal.

As testing of the prototype progressed, it was concluded that prohibiting all visual feedback was too restrictive. Audible feedback is at least 10 times slower than reading, which limits the amount of information that can be given to the user. Most of the time, data entry would progress with audible feedback alone. There will, however, be times when the user would be better served by looking up at a monitor to evaluate the state of the system, especially during error detection and resolution.

2.6.4 Conclusion

A prototype voice-driven data collection system for histopathology data using only voice input and computer-generated voice responses was developed and tested. Under controlled conditions, the overall accuracy rate was 97%. Additional work is needed to minimize training requirements and improve audible feedback. It was concluded that this architecture could be considered a viable alternative for data collection in animal toxicology studies with reasonable recognition accuracy. Two papers were published based on this work in Computers in Biology and Medicine [Grasso and Grasso 1994] and M.D. Computing [Grasso 1995].

3. Methodology

The general research hypothesis stated that speed, accuracy, and acceptance of a multimodal, multidimensional, human-computer interface will improve when the attributes are perceptually separable, and will improve for a unimodal interface when the attributes are perceptually integral. A set of software tools was developed to simulate a prototypical biomedical data collection task in order to test the validity of this hypothesis. The experiment was designed using repeated measures, with the order of conditions counterbalanced across all subjects. The following aspects of the experiment are discussed: independent variables, dependent variables, subjects, procedure, materials, analysis, and schedule.

3.1 Independent Variables

The two independent variables were interface (baseline, perceptually structured) and task order (slide group 1, slide group 2). The input task was to enter histopathologic observations consisting of three input attributes: topographical site, qualifier, and morphology. It was assumed that the qualifier/morphology (QM) relationship was integral, since the qualifier was 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 7 predicted which modality would lead to improvements in the computer-human 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 7: 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 8. 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. For testing, 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 8: Possible Interface Combinations for the Software Prototype

3.2 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 was recorded both by the experimenter and the software prototype. It was defined as the time it takes a participant to complete each of the 12 data entry tasks and was recorded to the nearest millisecond. The actual speed was determined by analysis of timing output from the prototype, recorded observations of the experimenter, and review of audio tapes recorded during the study.

Three measures of accuracy were recorded both by the experimenter and the software prototype: speech errors, mouse errors, and diagnosis errors. Speech recognition errors were counted when the prototype incorrectly recognized a spoken utterance by the participant. This was either because the participant was misunderstood by the prototype or the participant spoke a phrase that was not in 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 of a participant did not match the most likely diagnosis for each tissue slide. The actual 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 using 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 9 and the actual survey can be found in the Appendices in Section 6.3. 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.

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

Table 9: Adjective Pairs used in the User Acceptance Survey

3.3 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 was veterinary and clinical pathologists, graduate students and post-doctorates 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 Baltimore 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 shown in Table 10, based on responses to the pre-experiment questionnaire found in the Appendices in Section 6.1. The sample population consisted of professionals with advanced degrees, ranged in age from 33 to 51 years old, and were roughly equal in the number of males and females. Fifteen were from an academic institution, and most were U.S. born, native English speakers. The majority indicated they were comfortable using a computer with all but 3 ranking themselves with a 4 or higher in computer and mouse experience. Only 1 subject had any significant speech recognition experience.

Highest Degree	D.V.M.	11	Ph.D.	6	M.D.	3	
Institution	JHMI	8	UMMC	7	BVAMC	3	FDA 2
Age	Mean	40	Stdev	6.8			
Gender	Male	11	Female	9			
National Origin	US	13	Europe	4	India	2	Canada 1
Native Language	English	16	Other	4			
Computer Experience	Mean	5	Stdev	1.1			
Mouse Experience	Mean	5	Stdev	1.5			
Speech Experience	Mean	1	Stdev	0.9			

Table 10: Subject Demographics

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. The slides came from the National Center for Toxicological Research (Jefferson, AK). Two groups of six 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 11. 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. The actual slide diagnoses are shown in Table 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 11: Subject Groupings for the Experiment

Repeated measures (a within-groups design) is common among human computer interaction studies evaluating two or more input devices or other interface characteristics [Karl, Pettey, and Shneiderman 1992; Margono and Shneiderman 1993; Sears and Shneiderman 1991; Oviatt 1996]. Repeated observations on the same subject over time is a more efficient use of resources, since less participants are needed. Also, the estimation

of time trends is more precise, because measurements on the same subject tend to be less variable than measurements on different subjects [Keul 1994].

Group	*Slide*	*Diagnosis (Organ, Site, Qualifier, Morphology)*
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
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

Table 12: Tissue Slide Diagnoses

3.4 Procedure

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 found in the Appendices in Section 6.1. 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 head-set 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 actual 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 done to help ensure the experiment was measuring data input and not the ability of the subjects 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 Table 13.

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

Table 13: Experimental Procedure

3.5 Materials

A prototype computer program was developed for the experiment using Microsoft Windows 3.11 (Microsoft Corporation, Redmond, WA) and Borland C++ 4.51 (Borland International, Inc., Scotts Valley, CA). Some software components from the preliminary study described earlier were used in this effort. About 1,500 lines of code were written for two software programs. The first, ***pe_test***, supported the speech training task and the second, ***sm_test***, was used for the evaluation and data collection task.

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 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.

The main data entry task for the experiment was for subjects to view microscopic tissue slides and enter histopathologic observations. 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. The software provided prompts and directions to identify which modality was to be used for which inputs. A sample screen is shown in Figure 2.

The default operating mode for the PE500+ speech recognizer is called push-to-speak. In the push-to-speak mode, the user holds down a mouse button or foot pedal when speaking, so the recognizer knows when to process incoming utterances. The push-to-speak mode tended to have a higher recognition accuracy rate, but needed to be

avoided so as not to introduce additional effects into the experiment. Therefore, a software driver was developed that allowed the recognizer to operate in a voice-activated mode. Here, the PE500+ is always listening for speech input. Instead of having the user press and release a button, the start and end of an utterance was determined by signal amplitude levels, length of signal, and length of silence.

3.6 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). Testing each null hypothesis was done by computing the probability of obtaining that result. If the probability indicates that the result did not occur simply by chance, then the null hypothesis could be safely rejected [Johnson 1992].

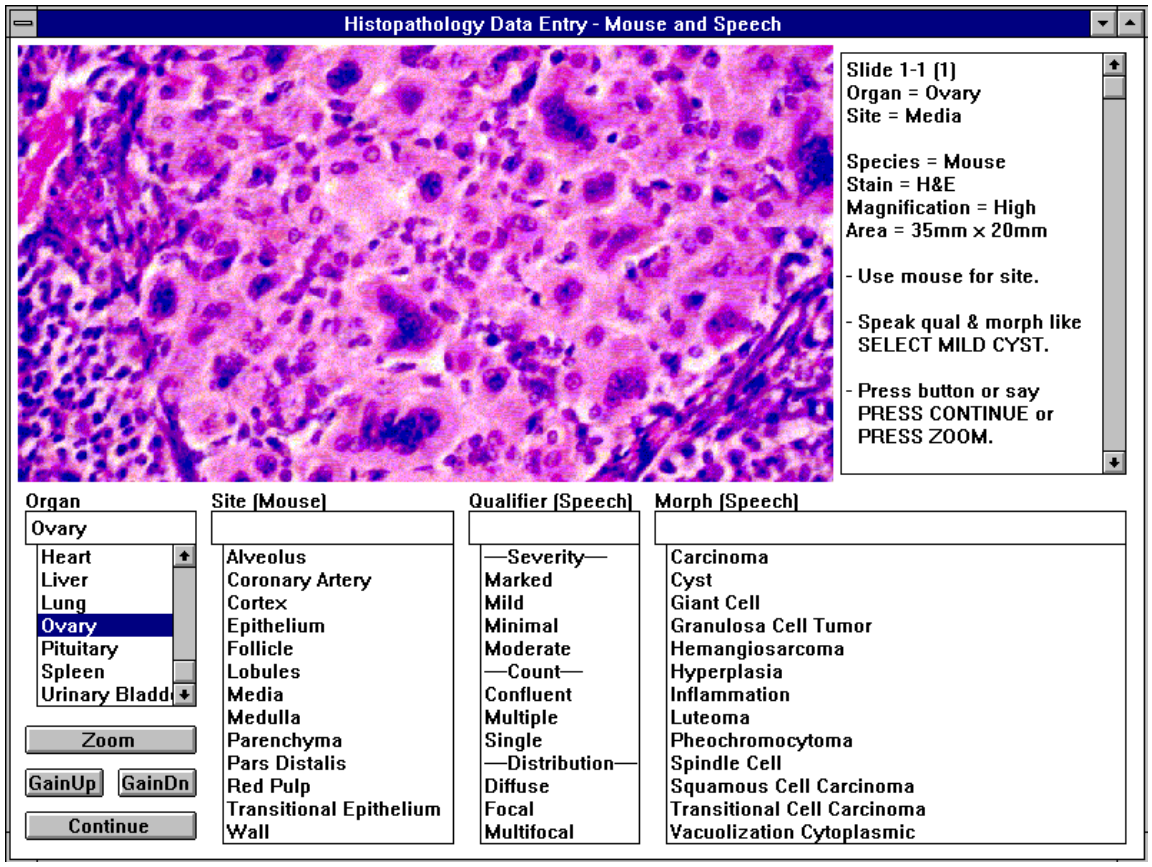

Figure 2: Sample Data Entry Screen

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.

3.7 Schedule and Deliverables

The success of these research objectives was demonstrated by completing and delivering the following items. The deliverables and work schedule are shown in Table 14 and Table 15. The schedule is based on a one-year effort, broken into three major parts: planning, operation, and interpretation [Basili 1986]. Included in experiment setup is a pilot study to evaluate the experimental procedures on a limited number of subjects. This allowed for changes to the experiment without biasing the pool of test subjects.

Task	***Duration***
Experiment design and software development	2 months
Pilot study	1 month
Retooling	1 month
Experiment operation	4 months
Analysis of results and publication development	4 months

Table 14: Research Schedule

Deliverables
The software prototype evaluated in the study
Data gathered from user testing (written, tape recorded, or video taped)
A Ph.D. Dissertation covering this effort in detail
One or more reports or publications based on this research

Table 15: Deliverables

4. Experimental Results

The experimental results include task completion times, speech errors, mouse errors, diagnosis errors, and the subjective questionnaire scores.

4.1 Task Completion Times

For each participant, a summary of the task completion times is shown in Table 16 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 group designation was described in Table 11. For example, B*1P2* means the subject used the baseline interface with slides 1 through 6 followed by the perceptually structured interface with slides 7 through 12. 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 as shown in Table 17 was significant ($F(1,38) = 4.719$, $p < .05$, two-tailed). A comparison of mean task completion times is in Figure 3 and a detailed listing of times is in the Appendices in Section 6.9.

Subject	*Group*	*Time for Baseline Tasks*	*Time for Perceptually Structured Tasks*	*Time Improvement*
1	B1P2	314.670	181.530	133.140
2	B2P1	195.230	147.770	46.460
3	P1B2	172.190	130.228	41.962
4	P2B1	122.537	96.888	25.649
5	B1P2	196.192	123.021	73.171
6	B2P1	120.725	106.499	14.226
7	P1B2	355.640	271.330	84.310
8	P2B1	185.867	127.708	58.159
9	B1P2	129.732	104.522	25.210
10	B2P1	159.777	134.786	24.991
11	P1B2	322.795	220.524	102.271
12	P2B1	128.140	103.809	24.331
13	B1P2	111.828	129.733	-17.905
14	B2P1	189.546	135.226	54.320
15	P1B2	153.241	132.205	21.036
16	P2B1	116.496	120.176	-3.680
17	B1P2	160.161	152.416	7.745
18	B2P1	209.695	133.907	75.788
19	P1B2	173.782	140.059	33.723
20	P2B1	169.341	165.892	3.449

Table 16: Times for the Baseline and Perceptually Structured Interfaces

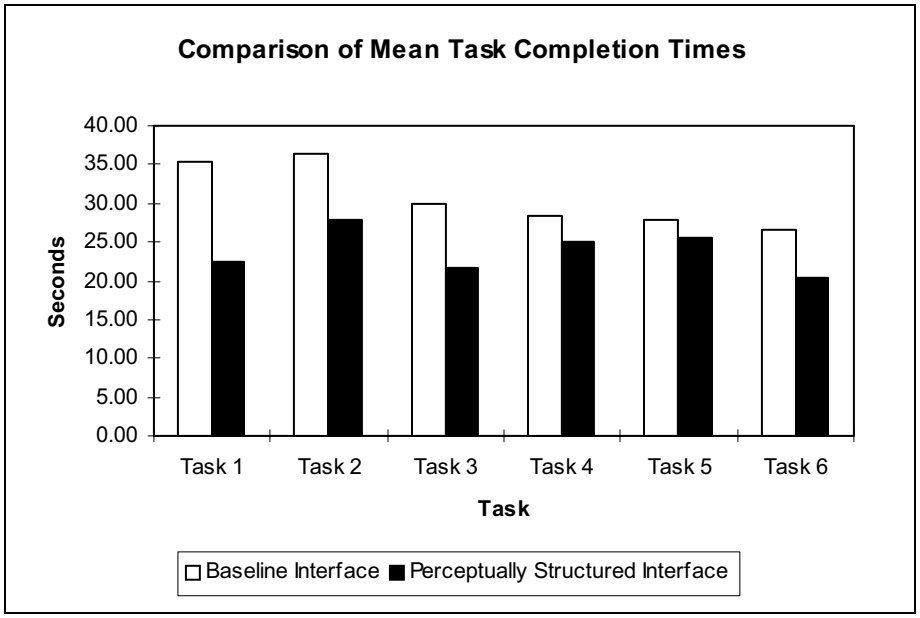

Figure 3: Comparison of Mean Task Completion Times

An analysis of variance (ANOVA) was performed to show that the interface order (baseline, perceptually structured) and task order (slide group 1, slide group 2) had no significant effect on the results. A single-factor ANOVA comparing the baseline-first-group and base-interface-second groups is shown in Table 18 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 shown in Table 19 was not significant ($F(1,18) = 0.723$, $p = 0.406$, two-tailed). A single factor ANOVA comparing the slide-group-one-first and slide-group-one-second groups shown in Table 20 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 shown in Table 21 was not significant ($F(1,18) = 1.650$, $p = 0.215$, two-tailed).

Single Factor ANOVA Comparing the Baseline and Perceptually Structured Interface Groups

Group	*Count*	*Sum*	*Mean*	*Variance*
Baseline	20	3687.585	184.379	4891.456
Perceptually Structured	20	2858.229	142.911	1731.705

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Value*	*P Value*
Between Groups	17195.784	1	17195.784	5.193	0.028
Within Groups	125840.054	38	3311.580		
Total	143035.838	39			

Table 17: ANOVA for Baseline and Perceptually Structured Interfaces

Single Factor ANOVA for the Baseline Interface Group

Group	*Count*	*Sum*	*Mean*	*Variance*
Baseline-First	10	1787.556	178.756	3453.106
Baseline-Second	10	1900.029	190.003	6803.022

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Value*	*P Value*
Between Groups	632.509	1	632.509	0.123	0.730
Within Groups	92305.146	18	5128.064		
Total	92937.655	19			

Table 18: Single Factor ANOVA for Baseline Groups

Single Factor ANOVA for the Perceptually Structured Interface Group

Group	*Count*	*Sum*	*Mean*	*Variance*
Perceptually-Structured-First	10	1508.819	150.882	3009.633
Perceptually-Structured-Second	10	1349.410	134.941	505.016

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Value*	*P Value*
Between Groups	1270.561	1	1270.561	0.723	0.406
Within Groups	31631.837	18	1757.324		
Total	32902.399	19			

Table 19: Single Factor ANOVA for Perceptually Structured Groups

Single Factor ANOVA for Slide Group 1

Group	*Count*	*Sum*	*Mean*	*Variance*
Slide-Group-One-First	10	1806.929	180.693	4700.170
Slide-Group-One-Second	10	1380.569	138.057	577.196

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Value*	*P Value*
Between Groups	9089.142	1	9089.142	3.445	0.080
Within Groups	47496.294	18	2638.683		
Total	56585.436	19			

Table 20: Single Factor ANOVA for Slide Group 1 Groups

Single Factor ANOVA for Slide Group 2

Group	*Count*	*Sum*	*Mean*	*Variance*
Slide-Group-Two-First	10	1489.446	148.945	1634.229
Slide-Group-Two-Second	10	1868.870	186.887	7090.527

Source of Variation	*Sum of Squares*	*Degrees of Freedom*	*Mean Square*	*F Value*	*P Value*
Between Groups	7198.129	1	7198.129	1.650	0.215
Within Groups	78522.803	18	4362.378		
Total	85720.932	19			

Table 21: Single Factor ANOVA for Slide Group 2 Groups

4.2 Errors

Three types of user errors were recorded: speech recognition errors, mouse errors, and diagnosis errors. A summary of error rates for each participant is shown in Table 22. A detailed listing of errors is in the Appendices in Section 6.10, Section 6.11, and Section 6.12. For speech errors, the baseline interface had mean 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). For mouse errors, the baseline interface had mean of 0.35 and 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 of 1.95 and the perceptually structured interface had mean of 1.90. Again, although the rate for the perceptually structured 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 4.

		--- Speech Errors ---		--- Mouse Errors ---		-- Diagnosis Errors -	
Subject	Group	Baseline	Perceptually Structured	Baseline	Perceptually Structured	Baseline	Perceptually Structured
1	B1P2	8	1	1	0	2	1
2	B2P1	7	6	0	0	2	3
3	P1B2	7	1	2	0	2	1
4	P2B1	3	1	0	0	0	0
5	B1P2	2	4	0	0	3	3
6	B2P1	8	2	0	0	2	4
7	P1B2	10	4	0	0	6	4
8	P2B1	8	4	0	0	4	5
9	B1P2	2	1	0	1	0	0
10	B2P1	2	2	0	0	2	2
11	P1B2	7	8	0	0	1	3
12	P2B1	4	3	2	0	1	0
13	B1P2	6	9	1	5	1	3
14	B2P1	4	2	0	0	1	1
15	P1B2	5	6	0	0	4	3
16	P2B1	3	1	1	0	1	2
17	B1P2	4	2	0	0	1	0
18	B2P1	7	4	0	0	2	0
19	P1B2	6	2	0	1	2	2
20	P2B1	5	3	0	0	3	1

Table 22: Baseline and Perceptually Structured Error Rates

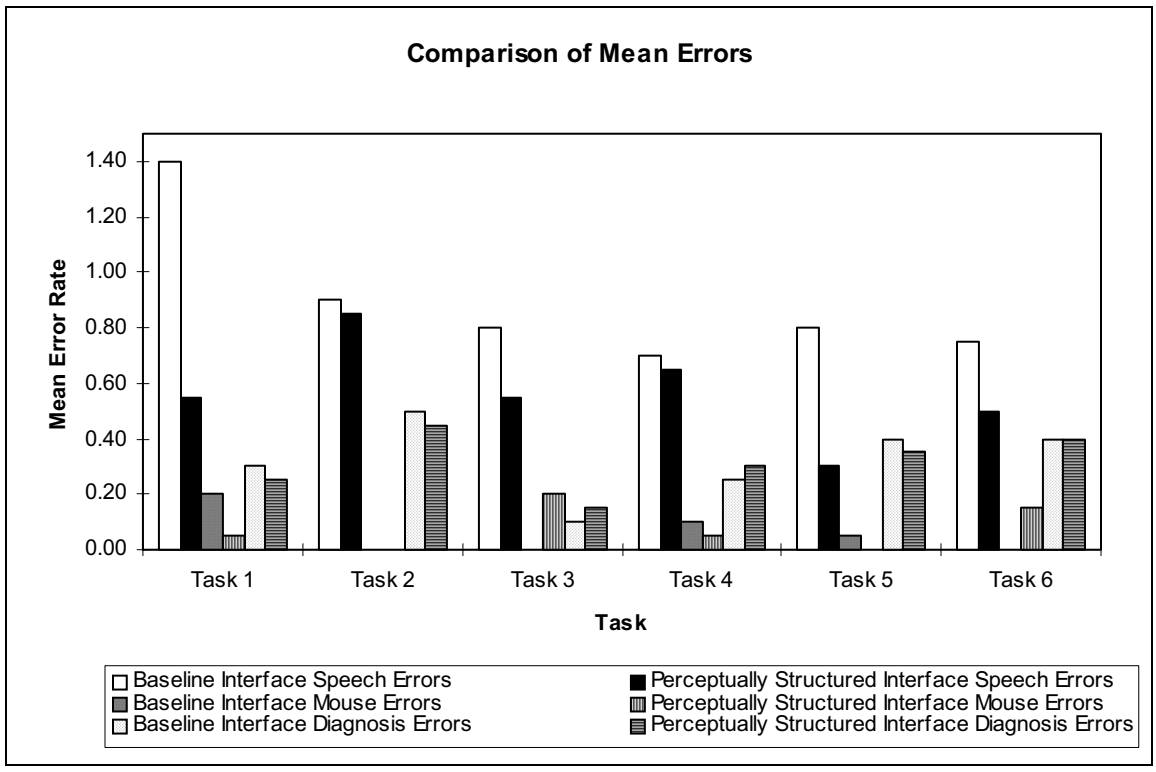

Figure 4: Comparison of Mean Errors

4.3 Acceptability

For analyzing the subjective scores, an acceptability index (AI) 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 as shown in Table 23 (p = .014). A comparison of these values is shown in Figure 5 and a summary of all acceptability scores is in the Appendices in Section 6.9.

Two-Factor ANOVA With Replication for Acceptability Index

SUMMARY	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Total
Baseline Interface														
Count	19	19	19	19	19	19	19	19	19	19	19	19	19	247
Sum	75	74	68	84	80	91	91	76	71	73	57	70	75	985
Average	3.95	3.89	3.58	4.42	4.21	4.79	4.79	4.00	3.74	3.84	3.00	3.68	3.95	3.99
Variance	3.27	1.99	2.48	3.37	2.51	2.73	1.95	2.11	2.20	2.70	2.89	4.12	2.27	2.75
Perceptually Structured Interface														
Count	19	19	19	19	19	19	19	19	19	19	19	19	19	247
Sum	62	66	66	75	78	82	85	74	57	70	48	66	67	896
Average	3.26	3.47	3.47	3.95	4.11	4.32	4.47	3.89	3.00	3.68	2.53	3.47	3.53	3.63
Variance	3.20	2.26	2.15	3.50	2.88	2.23	2.82	2.54	2.67	2.67	1.60	3.49	2.26	2.77
Total														
Count	38	38	38	38	38	38	38	38	38	38	38	38	38	
Sum	137	140	134	159	158	173	176	150	128	143	105	136	142	
Average	3.61	3.68	3.53	4.18	4.16	4.55	4.63	3.95	3.37	3.76	2.76	3.58	3.74	
Variance	3.27	2.11	2.26	3.40	2.62	2.47	2.35	2.27	2.51	2.62	2.24	3.71	2.25	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-Value*	*F Critical*
Sample	16.034	1	16.034	6.054	0.014	3.861
Columns	113.862	12	9.489	3.582	0.000	1.773
Interaction	5.255	12	0.438	0.165	0.999	1.773
Within	1239.579	468	2.649			
Total	1374.731	493				

Table 23: Two-Factor ANOVA for AI

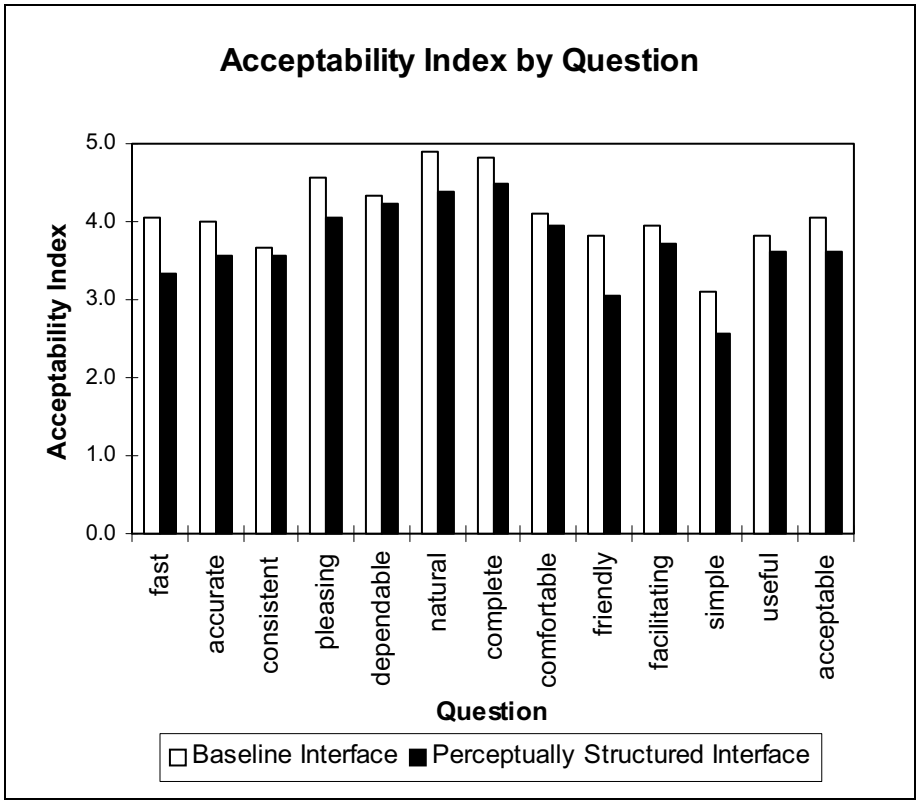

Figure 5: Comparison of Acceptability Index by Question

4.4 Correlation

The relationship between the dependent variables was analyzed using the Pearson correlation coefficient. These are time (T), speech errors (SE), mouse errors (ME), diagnosis errors (DE), and acceptability index (AI) from the baseline group, perceptually structured group, and both groups together. A summary of these coefficients is in Table 24. Representative graphs are shown in Figure 6, Figure 7, and Figure 8.

Variables	Sample Size	r value	Significant p value
Baseline Interface x Perceptually Structured Interface			
Baseline T x Perceptually Structured T	20	0.893	p < .001, two-tailed
Baseline SE x Perceptually Structured SE	20	0.223	
Baseline ME x Perceptually Structured ME	20	0.122	
Baseline DE x Perceptually Structured DE	20	0.667	p < .001, two-tailed
Baseline AI x Perceptually Structured AI	20	0.678	p < .001, two-tailed
T x SE			
Baseline T x Baseline SE	20	0.322	
Perceptually Structured T x Perceptually Structured SE	20	0.536	p < .05, two-tailed
T x SE	40	0.471	p < .01, two-tailed
T Improvement x Total SE	20	0.339	
T x ME			
Baseline T x Baseline ME	20	0.163	
Perceptually Structured T x Perceptually Structured ME	20	0.641	p < .01, two-tailed
T x ME	40	0.313	p < .05, two-tailed
T Improvement x Total ME	20	0.225	
T x DE			
Baseline T x Baseline DE	20	0.082	
Perceptually Structured T x Perceptually Structured DE	20	0.228	
T x DE	40	0.131	
T Improvement x Total DE	20	0.091	
T x AI			
Baseline T x Baseline AI	20	-0.120	
Perceptually Structured T x Perceptually Structured AI	20	0.018	
T x AI	40	-0.021	
T Improvement x Total AI	20	-0.134	
AI x SE			
Baseline AI x Baseline SE	20	0.264	
Perceptually Structured AI x Perceptually Structured SE	20	0.353	
AI x SE	40	0.324	p < .05, two-tailed
Total AI x Total SE	20	0.543	p < .05, two-tailed
AI x ME			
Baseline AI x Baseline ME	20	-0.489	p < .05, two-tailed
Perceptually Structured AI x Perceptually Structured ME	20	-0.039	
AI x ME	40	-0.187	
Total AI x Total ME	20	-0.237	
AI x DE			
Baseline AI x Baseline DE	20	0.425	p < .05, two-tailed
Perceptually Structured AI x Perceptually Structured DE	20	0.394	
AI x DE	40	0.407	p < .01, two-tailed
Total AI x Total DE	20	0.419	

Table 24: Pearson Correlation Coefficients for Dependent Variables

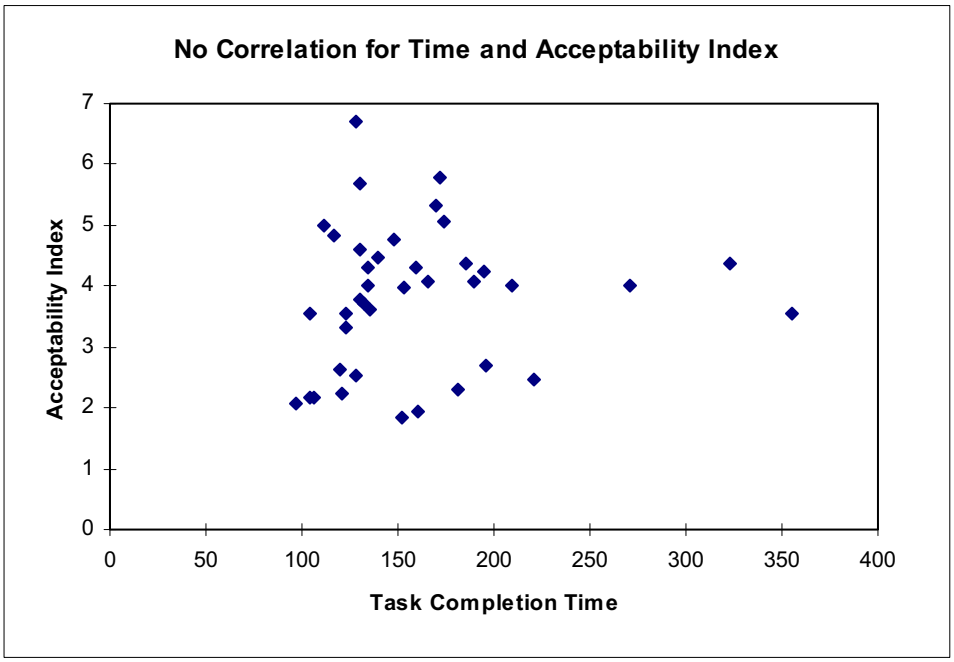

Figure 6: No Correlation Between Time and Acceptability Index

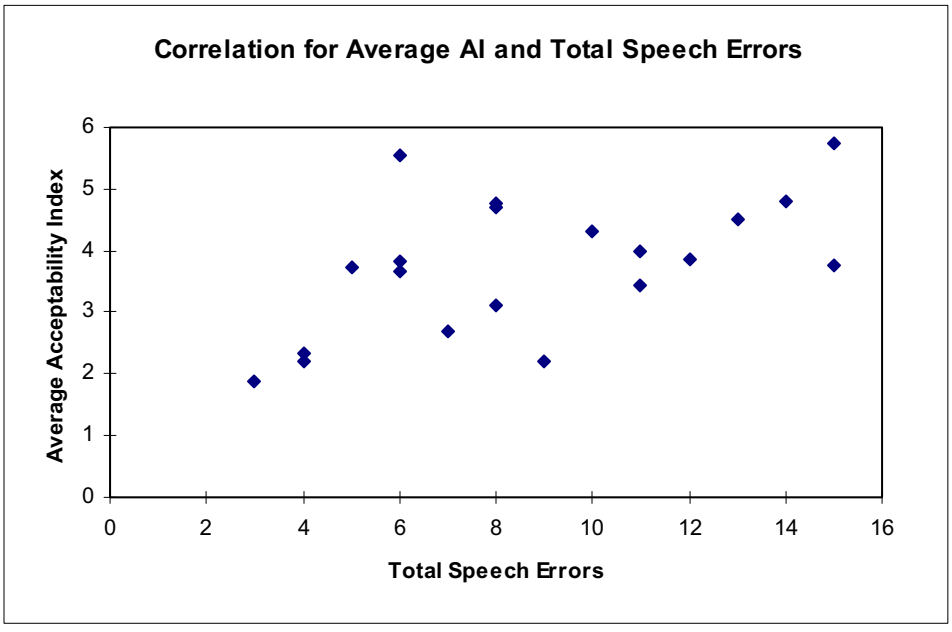

Figure 7: Correlation Between Average AI and Total Speech Errors

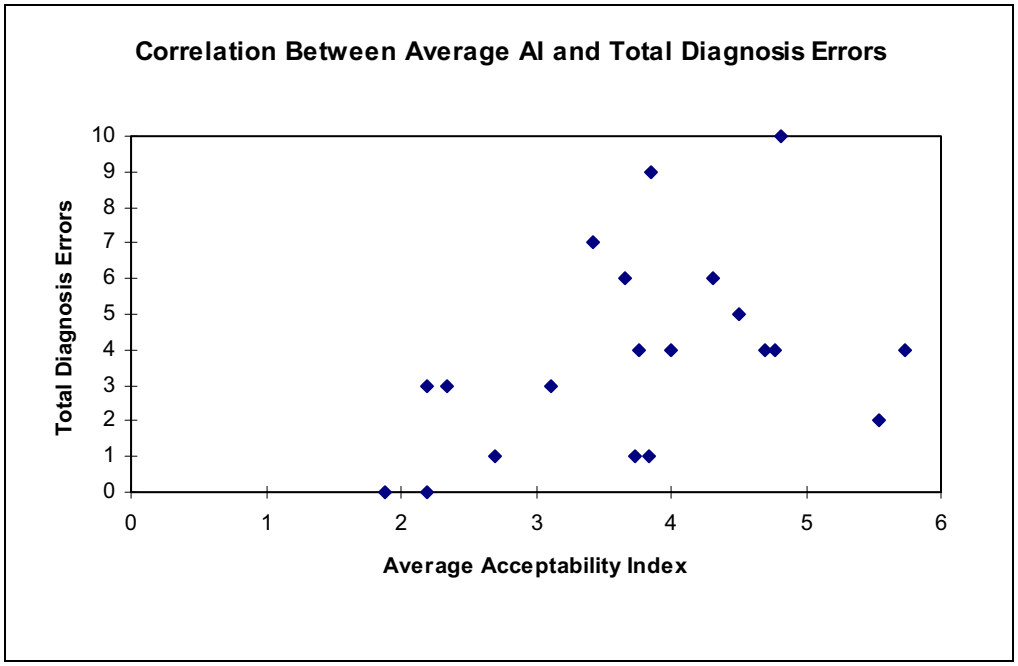

Figure 8: Correlation Between Average AI and Total Diagnosis Errors

5. Discussion and Conclusion

The results of this experiment support the hypothesis that the perceptual structure of an input task is an important consideration when designing multimodal computer interfaces. For multimodal speech and direct manipulation biomedical interfaces, the speed, accuracy, and acceptance of multidimensional input tasks improved when the attributes were perceived as separable. For unimodal interfaces, speed, accuracy, and acceptance improved when the inputs were perceived as integral. This chapter reviews the research findings, identifies possible relationships, summarizes the results, and outlines future research directions.

5.1 Findings

Three null hypotheses were identified before the study began. Two of the null hypotheses were rejected in favor of predicted results. One of the null hypotheses was rejected in part, in favor of predicted results.

The first null hypothesis stated: **($H1_0$)** *The integrality of input attributes has no effect on the speed of the user*. As reported in Section 4.1, 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% ($t(19) = 4.791$, $p < .001$, two-tailed). 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 (baseline, perceptually structured) and task order (slide group 1, slide group 2) had no significant effect on the results. The null hypothesis was rejected in support of an alternate hypothesis based on the predicted results: **($H1_A$)** *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 in Section 4.2, there were 1.95 less speech errors with the perceptually structured group, or a 36% improvement, with 16 of the 20 subjects having less errors using the perceptually structured interface. The reduction in speech errors was significant (paired $t(19) = 2.924$, $p < .01$, two-tailed). Mouse errors were slightly lower with the baseline group and diagnosis errors were slightly lower with the perceptually structured 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 a subject's expertise in identifying tissue types and reactions. The findings 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, since efforts were made to avoid measuring a subject's ability to apply domain expertise by allowing subjects 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 tasks 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 reported in Section 4.3, once the outlier was removed, the overall AI was 3.97 for the baseline group and 3.70 for the perceptually structured group. This was a moderate improvement of 6.7%, which was significant (2x13 ANOVA, $p < .05$). The null hypothesis was rejected in support of predicted results based on the alternate hypothesis: **(H3$_A$)** *The acceptance of multidimensional, multimodal interface will increase when the attributes of the task are perceived as*

separable, and for unimodal tasks will increase when the attributes of the task are perceived as integral.

One difficult aspect of collecting subjective data on user acceptance was that the prototype being tested was not a complete system. Subjects could view tissue slides on the screen, but were limited in other ways. The prototype allowed only one visual plane of the original slide to be examined, while pathologists typically require four such images to make a diagnosis. The zoom feature was limited. Also the 800x600 TFT panel was not the ideal computer monitor for viewing detail required to make diagnoses. A more complete system was developed as described under Preliminary Work in Section 2.6. While adding these and other features to this system was possible, they might have interfered with the independent variables being manipulated in the actual experiment. Nevertheless, it seemed difficult for some to subjectively evaluate a software prototype with limited functionality.

Another difficulty was with speech recognition accuracy. From informal testing during software development, the PE500+ accuracy rate was greater than 95% using push-to-speak mode but only about 80% using voice-activated mode. During the actual experiment, accuracy was 53% for the baseline interface and 64% for the perceptually structured interface. As described earlier, the voice-activated mode was used to avoid unwanted side effects when pressing a button to speak. However, this decreased accuracy frustrated some of the subjects, one of which compared it to yelling at your three-year-old: it doesn’t always work.

The perceptual structure of the input attributes used in this experiment might have been more subjective than originally anticipated. While most subjects who stated a preference selected the perceptually structured interface, some selected the baseline interface. In written comments, they viewed the morphology as the main term with the site and qualifier both modifying it. Using these assumptions, the baseline interface actually becomes more perceptually structured, since it uses separate modalities for the QM and SM input tasks and a single modality for SQ.

5.2 Relationships

The Pearson correlation coefficients, shown in Table 24, reveal possible relationships between the dependent variables. The following discussion examines why such relationships may exist.

5.2.1 Baseline Interface versus Perceptually Structured Interface

The positive correlation of time between the baseline interface and perceptually structured interface was anticipated. It was probably due to the fact that a subject who works slowly (or quickly) will do so regardless of the interface. 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 affected by the interface. Again, this was probably due to the fact that subjects were allowed to review the slides

before the actual test. The lack of correlation for mouse errors makes sense, since very few mouse errors were recorded.

The lack of correlation for speech errors was notable. If there was a positive correlation, it would imply 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, where the interface did not match the perceptual structure of the input task.

5.2.2 Relationships to Task Completion Time

One would expect that an increase in speech errors would result in an increase in task completion time, since it takes time to correct errors. Two of the coefficients in this group showed a positive correlation that was significant. They were time verses speech errors for the perceptually structured interface and time versus speech errors for both interfaces. The other two showed a positive correlation that was not significant, but was close.

Again, one would expect that an increase in mouse errors would result in an increase in task completion time. Two of the coefficients in this group did show a significant positive correlation and two did not. 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.

No correlation was observed between task completion time and the acceptability index. This result was similar to what was observed by Dillon [1995], who saw no correlation between time and acceptance, except with expert users. However, unlike Dillon, additional analysis found no correlation between time and acceptance with expert users. This was not necessarily a contradiction, because these two studies identified experts in different ways. Dillon identified a subject as an expert or novice based on a retrospective review of that person's work experience and education. Here, expertise was an independent variable. In contrast to that approach, this dissertation considered expertise a dependent variable and measured it prospectively, where expertise was inversely proportionate to the number of domain errors observed during the experiment.

5.2.3 Relationships with Acceptability Index

Between the acceptability index and speech errors, a significant positive correlation was observed for two of the four groups. This suggests than an increase in speech errors increases the likelihood the user will not be pleased with the interface. No correlation was found between the acceptability index and mouse errors. Again, this was

most likely due to the lack of recorded mouse errors. Note that for the acceptability index, a lower score corresponds to higher user acceptance.

A significant positive correlation was observed between the acceptability index and diagnosis errors. Three of the four showed this correlation, with the fourth being close. 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.

5.3 Summary

A research hypothesis was proposed for multimodal speech and direct manipulation biomedical 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 dissertation 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 natural language 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 errors decreased slightly and diagnosis errors increased slightly for the baseline interface, but these were not statistically significant. There was no relationship between user acceptance and time, suggesting that speed is not the predominate factor in determining approval. User acceptance was shown to be related to speech recognition errors, suggesting that recognition accuracy is crucial to user satisfaction. User acceptance was also shown to be related to domain errors, suggesting that the more domain expertise a person has, the more he or she will embrace the computer interface.

5.4 Future Research Directions

With respect to future directions, additional studies on domain expertise and minimizing speech errors would be helpful. This effort successfully reduced the rate of speech errors by applying certain 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 may be helpful.

As outlined under Research Questions, in Section 1.5, other experimental studies were proposed to further evaluate the framework for complementary behavior between speech and direct manipulation. This includes studies on the effect of the reference number, reference predictability, and reference visibility on the speed, accuracy, and acceptance of speech and direct manipulation interfaces. Additional research on speech input in multimodal environments, like this study, would also be of interest.

Preliminary work, described in Section 2.6, listed several areas of future research for speech interfaces in hands-busy, eyes-busy biomedical environments. Some of these, such as reducing speech training requirements, are being addressed by new technology. A key area warranting further study is how to improve audible feedback in eyes-busy tasks to reduce dependence on visual displays. A possible research goal could be to develop a fully functional speech-driven system incorporating results from the preliminary study and this dissertation that can be evaluated in production environments.

5.5 Conclusion

In conclusion, this study demonstrated that matching a multidimensional multimodal interface to the perceptual structure of the input attributes can increase the performance, accuracy, and user acceptance of the interface. User acceptance was influenced more by accuracy than speed. In addition, factors unrelated to the software itself affected acceptance, such as the level of domain expertise. It is hoped that these empirical results add to our understanding of how best to incorporate speech into multimodal environments and help in the development of systems to collect and manage biomedical information.

6. Appendices

Memoranda, questionnaires, vocabulary, transcripts, and experimental data are included here.

6.1 Sample Memorandum to Request for Volunteers

To: Veterinary Pathologists

From: Michael Grasso
University of Maryland Baltimore County
grasso@cs.umbc.edu

Date: December 10, 1996

Subj: Volunteers Needed for Biomedical Software Study

I am a doctoral student in Computer Science at the University of Maryland Baltimore County. My dissertation centers on the acceptance and efficiency of computers using a spoken language interface in biomedical environments.

Part of my research is to evaluate several user interfaces in the context of collecting histopathology data. I need up to 40 participants who can volunteer about 45 minutes of their time between now and March 1997. The participants should be clinical or veterinary pathologists, graduate students, residents, or post-doctorates who feel comfortable with tissue types and reactions. Note that since the main objective is to evaluate different user interfaces, participants do not need a high level of expertise in this area.

The test is relatively simple and lasts only about 45 minutes. Each participant will be asked to enter histopathologic observations as the software projects images of tissue slides on a computer monitor. Testing will take place at your facility, so not travel is involved.

I'll call in the near future to see if you might be able to help. In the meantime, if you have questions, you can reach me by e-mail (grasso@cs.umbc.edu, 410-455-3000). Also, feel free to talk my advisors at UMBC, Dr. Finin (finin@cs.umbc.edu, 410-455-3522) and Dr. Ebert (ebert@cs.umbc.edu, 410-455-3541).

Thanks for taking the time to consider this. Your assistance in this project will be greatly appreciated.

6.2 Pre-Experiment Questionnaire

Pre-Experiment Questionnaire Subject # ____

Answer the following questions and rate your level of computer experience by circling the appropriate number for the scaled items below.

1. Occupation: _____

2. Education (Highest Degree and Major) :_____

3. Age:_____ 4. Gender: _____

5. National Origin: _____

6. Native Language: _____

Level of Computer Experience

7. none high proficiency
   1 2 3 4 5 6 7

Level of Experience using a Computer with a Mouse

8. none high proficiency
   1 2 3 4 5 6 7

Level of Experience using a Computer with Speech Recognition

9. none high proficiency
   1 2 3 4 5 6 7

Informed Consent

I have read the above instructions, and I have answered all questions to my satisfaction. I understand that my responses may be recorded and that only the investigator and I will have access to these results by name. I agree to participate in this research project.

Name:_____ Signature: _____

Date: _____ Investigator: _____

6.3 Post-Experiment Questionnaire

Post-Experiment Questionnaire **Subject # ____ Group ____**

For each interface, rate your satisfaction by circling the appropriate number for the scaled items below. Select number 4 if neutral.

	Interface 1							**Interface 2**						
1)	fast						slow	fast						slow
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
2)	accurate					inaccurate		accurate					inaccurate	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
3)	consistent					inconsistent		consistent					inconsistent	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
4)	pleasing					irritating		pleasing					irritating	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
5)	dependable				undependable			dependable				undependable		
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
6)	natural					unnatural		natural					unnatural	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
7)	complete					incomplete		complete					incomplete	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
8)	comfortable				uncomfortable			comfortable				uncomfortable		
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
9)	friendly					unfriendly		friendly					unfriendly	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
10)	facilitating					distracting		facilitating					distracting	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
11)	simple					complicated		simple					complicated	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
12)	useful					useless		useful					useless	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Overall Evaluation:														
13)	acceptable					unacceptable		acceptable					unacceptable	
	1	2	3	4	5	6	7	1	2	3	4	5	6	7

Comments: If you like, you can write additional comments on the back of this form.

6.4 Pathology Nomenclature

Organs
Adrenal Gland
Heart
Liver
Lung
Ovary
Pituitary
Spleen
Urinary Bladder

Sites
Alveolus
Coronary Artery
Cortex
Epithelium
Follicle
Lobules
Media
Medulla
Parenchyma
Pars Distalis
Red Pulp
Transitional Epithelium
Wall

Severity Qualifiers
Marked
Mild
Minimal
Moderate

Count Qualifiers
Confluent
Multiple
Single

Distribution Qualifiers
Diffuse
Focal
Multifocal

Morphologies
Carcinoma
Cyst
Giant Cell
Granulosa Cell Tumor
Hemangiosarcoma
Hematopoietic Cell Proliferation Erythrocytic
Hepatocellular Carcinoma
Hyperplasia
Inflammation
Luteoma
Pheochromocytoma
Squamous Cell Carcinoma
Spindle Cell
Transitional Cell Carcinoma
Vacuolization Cytoplasmic

6.5 Perceptually Structured Interface Vocabulary

S -> { Select OBS+ | Press Continue | Press Zoom }

OBS+ -> QUAL_LIST+ MORPH_LIST+

QUAL_LIST+ == Marked Mild Minimal Moderate
Confluent Multiple Single
Diffuse Focal Multifocal

MORPH_LIST+ -> {Carcinoma | Cyst | Giant Cell | Granulosa Cell Tumor |
Hemangiosarcoma |
Hematopoietic Cell Proliferation Erythrocytic |
Hepatocellular Carcinoma | Hyperplasia |
Inflammation | Luteoma | Pheochromocytoma |
Squamous Cell Carcinoma | Spindle Cell |
Transitional Cell Carcinoma |
Vacuolization Cytoplasmic}

6.6 Baseline Interface Vocabulary

S -> { Select OBS+ | Press Continue | Press Zoom }

OBS+ -> SITE_LIST+ QUAL_LIST+

SITE_LIST+ -> {Alveolus | Coronary Artery | Cortex | Epithelium |
Follicle | Lobules | Media | Medulla | Parenchyma |
Pars Distalis | Red Pulp | Transitional Epithelium |
Wall}

QUAL_LIST+ == Marked Mild Minimal Moderate
Confluent Multiple Single
Diffuse Focal Multifocal

6.7 Perceptually Structured Interface Transcript

Following is a transcript from subject 12 using the perceptually structured interface showing elapsed time in seconds along with the device used, the action, and comments.

Time	***Device***	***Action***	***Comment***
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	
20	Mouse	Click on “follicle”	
29	Speech	“Select moderate hyperplasia”	Recognition error
36	Speech	“Select moderate hyperplasia”	
42	Mouse	Click on “press continue” button	
44	Mouse	Click on “media”	
50	Speech	“Select moderate inflammation”	
57	Mouse	Click on “press continue” button	
61	Mouse	Click on “wall”	
65	Speech	“Select marked squamous cell carcinoma”	
71	Mouse	Click on “press continue” button	
74	Mouse	Click on “epithelium”	
81	Speech	“Select moderate transitional cell carcinoma”	
89	Mouse	Click on “press continue” button	
94	Mouse	Click on “transitional epithelium”	
96	Speech	“Select marked transitional cell carcinoma”	
104	Mouse	Click on “press continue” button	

6.8 Baseline Interface Transcript

Following is a transcript from subject 12 using the baseline interface showing elapsed time in seconds along with the device used, the action, and comments.

Time	***Device***	***Action***	***Comment***
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	
35	Speech	“Select cortex marked”	Recognition error
39	Mouse	Click on “pheochromocytoma”	
42	Speech	“Select cortex marked”	
51	Mouse	Click on “press continue” button	
70	Speech	“Select pars distalis moderate”	
76	Mouse	Click on “granulosa cell tumor”	
77	Mouse	Click on “press continue” button	
82	Speech	“Select lobules marked”	
88	Mouse	Click on “vacuolization cytoplasmic”	
89	Mouse	Click on “press continue” button	
97	Speech	“Select parenchyma moderate”	Recognition error
101	Mouse	Click on “hemangiosarcoma”	
103	Speech	“Select parenchyma moderate”	
109	Mouse	Click on “press continue” button	
114	Speech	“Select parenchyma marked”	Recognition error
118	Mouse	Click on “hepatocellular carcinoma”	
124	Speech	“Select parenchyma marked”	
128	Mouse	Click on “press continue” button	

6.9 Task Completion Time Scores

	Subject	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total	Improvement
Baseline	1	22.410	72.170	50.090	49.490	58.000	62.510	314.670	
Interface	2	28.080	16.300	62.740	37.960	30.210	19.940	195.230	
	3	54.650	48.279	32.571	20.267	21.750	18.675	196.192	
	4	10.051	16.203	15.983	30.813	24.277	23.398	120.725	
	5	11.808	21.476	23.124	25.375	24.332	23.617	129.732	
	6	49.323	25.320	25.925	26.309	15.983	16.917	159.777	
	7	17.851	18.510	18.509	20.487	20.762	15.709	111.828	
	8	56.463	15.928	18.564	20.872	26.639	51.080	189.546	
	9	32.626	43.775	26.309	18.510	25.485	13.456	160.161	
	10	28.213	69.550	45.097	20.845	18.520	27.470	209.695	
	11	25.046	33.669	29.824	20.926	24.552	38.173	172.190	
	12	11.973	20.432	13.786	14.061	24.497	37.788	122.537	
	13	76.621	103.643	63.109	58.605	35.976	17.686	355.640	
	14	17.576	52.673	28.946	28.89	38.997	18.785	185.867	
	15	68.162	64.702	54.156	28.616	66.514	40.645	322.795	
	16	27.737	23.453	25.925	12.742	18.675	19.608	128.140	
	17	43.995	21.311	13.292	37.568	23.014	14.061	153.241	
	18	22.519	14.171	18.4	20.871	13.182	27.353	116.496	
	19	50.476	30.483	14.61	41.633	16.862	19.718	173.782	
	20	49.539	15.189	17.966	31.806	29.067	25.774	169.341	
Perceptually	1	24.06	37.74	21.03	40.15	36.47	22.08	181.530	133.140
Structured	2	10.93	22.68	34.88	32.9	23.34	23.04	147.770	47.460
Interface	3	17.071	15.873	16.807	17.741	35.042	20.487	123.021	73.171
	4	13.292	26.529	17.026	22.19	16.313	11.149	106.499	14.226
	5	12.852	14.006	22.245	15.104	15.269	25.046	104.522	25.210
	6	14.115	33.395	27.627	25.430	16.917	17.302	134.786	24.991
	7	40.645	12.029	18.070	26.419	17.081	15.489	129.733	-17.905
	8	18.016	8.184	12.248	55.859	15.159	25.760	135.226	54.320
	9	18.839	66.624	19.279	14.445	15.434	17.795	152.416	7.745
	10	26.068	18.801	21.060	27.194	21.477	19.307	133.907	75.788
	11	22.245	42.622	13.896	16.642	23.014	11.809	130.228	41.962
	12	11.754	15.709	28.945	13.732	14.280	12.468	96.888	25.649
	13	31.033	58.385	44.489	42.402	75.248	19.773	271.330	84.310
	14	22.355	14.225	13.457	12.578	24.936	40.157	127.708	58.159
	15	49.268	52.179	20.432	37.843	31.967	28.835	220.524	102.271
	16	14.446	28.396	13.731	14.665	17.960	14.611	103.809	24.331
	17	20.597	24.991	26.09	28.286	15.928	16.313	132.205	21.036
	18	23.783	15.159	21.805	19.993	18.400	21.036	120.176	-3.680
	19	29.275	26.804	22.079	15.160	19.223	27.518	140.059	33.723
	20	26.292	25.391	18.245	22.143	55.721	18.100	165.892	3.449

6.10 Speech Errors

	Subject	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total
Baseline	1	0	2	2	2	1	1	**8**
Interface	2	1	0	3	0	2	1	**7**
	3	5	2	0	0	0	0	**7**
	4	0	0	0	1	1	1	**3**
	5	0	0	0	0	1	1	**2**
	6	4	2	1	1	0	0	**8**
	7	1	2	2	1	2	2	**10**
	8	2	0	1	1	1	3	**8**
	9	2	0	0	0	0	0	**2**
	10	1	3	2	0	0	1	**7**
	11	3	3	0	0	1	0	**7**
	12	1	0	3	0	0	0	**4**
	13	0	2	1	2	1	0	**6**
	14	0	0	0	0	3	1	4
	15	4	0	1	0	0	0	**5**
	16	0	1	0	0	1	1	**3**
	17	0	0	0	3	1	0	**4**
	18	0	0	0	0	0	1	**1**
	19	2	1	0	2	0	1	**6**
	20	2	0	0	1	1	1	**5**
Total		**28**	**18**	**16**	**14**	**16**	**15**	**107**
Perceptually	1	1	0	0	0	0	0	**1**
Structured	2	0	1	3	2	0	0	**6**
Interface	3	0	0	0	0	1	0	**1**
	4	0	1	0	0	0	0	**1**
	5	1	0	1	1	0	1	**4**
	6	0	1	0	1	0	0	**2**
	7	2	0	1	0	1	0	**4**
	8	0	0	0	3	0	1	**4**
	9	0	1	0	0	0	0	**1**
	10	2	0	1	1	0	0	**4**
	11	2	3	0	0	1	2	**8**
	12	0	0	0	0	0	3	**3**
	13	0	0	3	4	0	2	**9**
	14	0	2	0	0	0	0	**2**
	15	1	4	0	0	1	0	**6**
	16	0	1	0	0	0	0	**1**
	17	0	1	0	1	0	0	**2**
	18	1	0	1	0	1	1	**4**
	19	0	1	1	0	0	0	**2**
	20	1	1	0	0	1	0	**3**
Total		**11**	**17**	**11**	**13**	**6**	**10**	**68**

6.11 Mouse Errors

	Subject	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total
Baseline	1	0	0	0	0	1	0	**1**
Interface	2	0	0	0	0	0	0	**0**
	3	2	0	0	0	0	0	**2**
	4	0	0	0	0	0	0	**0**
	5	0	0	0	0	0	0	**0**
	6	0	0	0	0	0	0	**0**
	7	0	0	0	0	0	0	**0**
	8	0	0	0	0	0	0	**0**
	9	0	0	0	0	0	0	**0**
	10	0	0	0	0	0	0	**0**
	11	0	0	0	0	0	0	**0**
	12	1	0	0	1	0	0	**2**
	13	0	0	0	1	0	0	**1**
	14	0	0	0	0	0	0	**0**
	15	0	0	0	0	0	0	**0**
	16	1	0	0	0	0	0	**1**
	17	0	0	0	0	0	0	**0**
	18	0	0	0	0	0	0	**0**
	19	0	0	0	0	0	0	**0**
	20	0	0	0	0	0	0	**0**
Total		**4**	**0**	**0**	**2**	**1**	**0**	**7**
Perceptually	1	0	0	0	0	0	0	**0**
Structured	2	0	0	0	0	0	0	**0**
Interface	3	0	0	0	0	0	0	**0**
	4	0	0	0	0	0	0	**0**
	5	0	0	0	0	0	0	**0**
	6	0	0	0	0	0	0	**0**
	7	0	0	0	0	0	0	**0**
	8	0	0	0	0	0	0	**0**
	9	0	0	1	0	0	0	**1**
	10	0	0	0	0	0	0	**0**
	11	0	0	0	0	0	0	**0**
	12	0	0	0	0	0	0	**0**
	13	0	0	3	0	0	2	**5**
	14	0	0	0	0	0	0	**0**
	15	0	0	0	0	0	0	**0**
	16	0	0	0	0	0	0	**0**
	17	0	0	0	0	0	0	**0**
	18	0	0	0	1	0	1	**2**
	19	1	0	0	0	0	0	**1**
	20	0	0	0	0	0	0	**0**
Total		**1**	**0**	**4**	**1**	**0**	**3**	**9**

6.12 Diagnosis Errors

	Subject	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total
Baseline	1	0	1	0	0	1	0	**2**
Interface	2	0	1	0	0	0	1	**2**
	3	1	0	0	0	0	1	**2**
	4	0	0	0	0	0	0	**0**
	5	0	1	1	0	1	0	**3**
	6	0	1	0	0	1	0	**2**
	7	1	1	1	1	1	1	**6**
	8	1	0	0	1	1	1	**4**
	9	0	0	0	0	0	0	**0**
	10	0	0	0	1	1	0	**2**
	11	0	0	0	1	0	0	**1**
	12	0	1	0	0	0	0	**1**
	13	0	0	0	0	1	0	**1**
	14	0	0	0	0	0	1	**1**
	15	1	1	0	0	1	1	**4**
	16	0	1	0	0	0	0	**1**
	17	0	0	0	1	0	0	**1**
	18	0	0	0	0	0	1	**1**
	19	1	1	0	0	0	0	**2**
	20	1	1	0	0	0	1	**3**
Total		**6**	**10**	**2**	**5**	**8**	**8**	**39**
Perceptually	1	0	1	0	0	0	0	**1**
Structured	2	1	0	1	0	1	0	**3**
Interface	3	0	1	0	0	0	0	**1**
	4	0	0	0	0	0	0	**0**
	5	0	1	0	0	1	1	**3**
	6	1	0	1	1	0	1	**4**
	7	0	0	1	1	1	1	**4**
	8	1	1	0	1	1	1	**5**
	9	0	0	0	0	0	0	**0**
	10	0	1	0	1	0	0	**2**
	11	0	0	0	1	1	1	**3**
	12	0	0	0	0	0	0	**0**
	13	1	1	0	0	0	1	**3**
	14	0	1	0	0	0	0	**1**
	15	1	1	0	0	1	0	**3**
	16	0	1	0	0	0	1	**2**
	17	0	0	0	0	0	0	**0**
	18	0	0	0	0	0	0	**0**
	19	0	0	0	1	1	0	**2**
	20	0	0	0	0	0	1	**1**
Total		**5**	**9**	**3**	**6**	**7**	**8**	**38**

6.13 Acceptability Scores

	Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	AI[subject]
Baseline	1	2	2	2	2	2	3	4	2	2	2	1	1	2	**2.08**
Interface	2	6	3	5	4	5	5	4	3	5	4	2	4	5	**4.23**
	3	3	3	2	3	3	3	3	3	3	3	2	2	2	**2.69**
	4	2	2	2	2	2	3	3	3	2	2	2	2	2	**2.23**
	5	3	4	5	4	5	3	5	3	3	3	5	3	3	**3.77**
	6	2	3	2	4	3	6	6	5	4	5	5	7	4	**4.31**
	7	4	5	5	6	5	7	7	6	4	6	2	4	4	**5.00**
	8	5	4	4	5	4	5	4	4	3	4	4	3	4	**4.08**
	9	2	3	1	1	1	3	5	2	2	1	1	1	2	**1.92**
	10	2	4	5	5	5	4	4	3	4	4	4	4	4	**4.00**
	11	7	6	6	7	6	6	7	6	4	6	2	6	6	**5.77**
	12	5	2	2	4	3	4	3	4	4	4	3	2	3	**3.31**
	13	3	3	2	6	6	3	4	5	2	3	2	4	3	**3.54**
	14	7	7	4	7	7	7	7	7	7	7	6	7	7	**6.69**
	15	6	5	5	5	5	4	5	4	6	2	2	3	5	**4.38**
	16	2	3	2	2	3	4	3	2	2	3	2	2	3	**2.54**
	17*	1	1	1	1	1	1	1	1	1	1	1	1	1	**1.00**
	18	6	5	5	4	5	7	5	4	6	3	6	2	5	**4.85**
	19	4	5	4	7	5	7	6	5	4	6	1	6	6	**5.08**
	20	4	5	5	6	5	7	6	5	4	5	5	7	5	**5.31**
AI[question]		**3.80**	**3.75**	**3.45**	**4.25**	**4.05**	**4.60**	**4.60**	**3.85**	**3.60**	**3.70**	**2.90**	**3.55**	**3.80**	**3.84**
Perceptually	1	2	2	2	2	2	3	4	3	2	3	2	1	2	**2.31**
Structured	2	5	6	5	4	6	5	5	4	5	4	3	5	5	**4.77**
Interface	3	7	3	3	6	3	3	7	2	3	3	2	2	2	**3.54**
	4	2	2	2	2	2	3	3	3	2	2	1	2	2	**2.15**
	5	3	4	3	3	4	5	5	3	3	4	3	3	3	**3.54**
	6	3	2	3	4	3	6	4	5	4	6	5	7	4	**4.31**
	7	1	4	4	6	6	7	7	7	3	4	2	4	5	**4.62**
	8	3	3	5	3	4	5	5	5	3	3	2	3	3	**3.62**
	9	2	3	2	2	2	2	3	2	1	1	1	1	2	**1.85**
	10	2	4	5	5	5	4	4	3	4	4	4	4	4	**4.00**
	11	6	6	6	7	6	6	7	6	4	6	2	6	6	**5.69**
	12	3	2	2	2	2	2	2	2	2	2	2	2	2	**2.08**
	13	4	4	3	6	6	4	4	5	2	4	2	4	4	**4.00**
	14	7	7	4	7	7	7	7	7	7	7	6	7	7	**6.69**
	15	2	2	2	2	3	4	3	4	1	2	2	2	3	**2.46**
	16	2	3	2	1	3	3	3	2	1	2	2	2	2	**2.15**
	17*	4	4	4	4	4	4	4	4	4	4	4	4	4	**4.00**
	18	2	3	2	4	3	4	2	3	1	3	2	2	3	**2.62**
	19	4	2	6	4	6	5	6	4	4	6	2	4	5	**4.46**
	20	2	4	5	5	5	4	4	4	5	4	3	5	3	**4.08**
AI[question]		**3.30**	**3.50**	**3.50**	**3.95**	**4.10**	**4.30**	**4.45**	**3.90**	**3.05**	**3.70**	**2.60**	**3.50**	**3.55**	**3.65**

* Note that subject 17 was considered an outlier and removed during analysis.

7. References

Ahlberg, C., Williamson, C., Shneiderman, B. (1992). Dynamic Queries for Information Exploration: An Implementation and Evaluation, ACM CHI '92 Conference Proceedings, pp. 619-626, Monterey, CA, May 3-7.

Basili, V. R., Welby, R. W., Hutchens, D. H. (1986). Experimentation in Software Engineering. IEEE Transactions on Software Engineering. 12(7):733-743.

Bergeron, B. and Locke, S. (1990). Speech Recognition as a User Interface. M.D. Computing, 7(5):329-334.

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.

Carbonell, N. (1994). Multimodal Human-Computer Interaction. ACM SIGCHI Bulletin, 26(3):15-18.

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.

CERN European Laboratory for Particle Physics. World-Wide Web Home, URL http://infor.cern.ch/, undated.

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.

Conte, E. (1994). A Basic HTML Style Guide: Readability. URL: http://heasarc.gsfc.nasa.gov /0/docs/heasarc/Style_Guide/readability.html.

Cranmer, M. F., Lawrence, L. R., Konvicka, A. J., Herrick, S. S. (1978). NCTR Computer Systems Designed for Toxicologic Experimentation. I. Overview. Journal of Environmental Pathology and Toxicology, 1(5):701-709.

Daly, A. M., Martin, R. A., McGuire, E. J., DiFonzo, C. J. (1989). A Microcomputer-Based Data Acquisition and Reporting System for Clinical Pathology Data from Animal Drug Toxicology Studies. Drug Information Journal, 23:285-296.

Damper, R. I. (1993). Speech as an Interface Medium: How can it Best be Used? In Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers, Taylor & Fancis.

Davis, H. K., Biddulph, R., Balashek, S. (1952). Automatic Recognition of Spoken Digits. American Journal of Otolaryngology, 24:637-642.

Dershaw, D. D. (1988). Voice-Activated Radiology Reports. Radiology, 187:284.

Dillon, T. W., Norcio, A. F., DeHaemer, M. J. (1993). Spoken Language Interaction: Effects of Vocabulary Size and Experience on User Efficiency and Acceptability. In Proceedings of the Fifth International Conference on Human-Computer Interaction, pp. 140-145, Orlando, Florida.

Dillon, T. W., McDowell, D., Norcio, A. F., DeHaemer, M. J. (1994). Nursing Acceptance of a Speech-Input Interface: A Preliminary Investigation. Computers in Nursing, 12(6):264-271.

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.

Dudley, H. Balashek, S. (1958). Automatic Recognition of Phonetic Patterns in Speech. Journal of the Acoustic Society of America, 30:721-739.

Faccini, J. M., Naylor, D. (1979). Computer Analysis and Integration of Animal Pathology Data. Archives of Toxicology, Supplement, 2:517-520.

Feldman, C. A., Stevens, D. (1990). Pilot Study on the Feasibility of a Computerized Speech Recognition Charting System. Community Dentistry and Oral Epidemiology, 18:213-215.

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

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., 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.

Green, G. (1993). Director of Strategic Systems Planning, U.S. Food and Drug Administration. Personal Interview.

Hollbrook, J. A. (1992). Generating Medical Documentation Through Voice Input: The Emergency Room. Topics in Health Records Management, 12(3):58-63.

House D. (1995). Spoken-Language Access to Multimedia (SLAM): A Multimodal Interface to the World-Wide Web. Masters Thesis, Oregon Graduate Institute.

Ikerira, H., et al. (1990). Analysis of Bone Scintigram Data Using Speech Recognition Reporting System. Radiation Medicine, 8(1):8-12.

Issacs, E., Wulfman, C. E., Rohn, J. A., Lane, C. D., Fagan, L. M. (1993). Graphical Access to Medical Expert System: IV. Experiments to Determine the Role of Spoken Input. Methods of Information in Medicine, 32(1):18-32.

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.

Karl, L., Pettey, M., Shneiderman, B. (1992). Speech-Activated versus Mouse-Activated Commands for Word Processing Applications: An Empirical Evaluation. Technical Report CAR-TR-630, Center for Automation Research, University of Maryland, College Park, MD.

Kassel, R. H. (1995). A Comparison of Approaches to On-Line Handwritten Character Recognition. Doctoral Dissertation, Massachusetts Institute of Technology.

Keul, R. (1994). Statistical Priciples of Research Design and Analysis. Duxbury Press, Belmont, CA, page 499.

Klatt, E. C. (1991). Voice-Activated Dictation for Autopsy Pathology. Computers in Biology and Medicine, 21(6):429-433.

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.

Lea, W. A. (1980). Trends in Speech Recognition, Prentice Hall, Englewood Cliffs, NJ.

Lefebvre, P., Duncan, G., and Poirier, F. (1993). Speaking with Computers: A Multimodal Approach. In Proceedings of EuroSpeech ‘93, pp. 1665-1669, Berlin.

Maberly, N.C. (1966). Mastering Speed Reading, New American Library, Inc., New York.

Margono, S. and Shneiderman, B. (1993). A Study of File Manipulation by Novices using Commands vs. Direct Manipulation. In Sparks of Innovation in Human-Computer Interaction, Ablex Publishing Corporation, Norwood, NJ.

Massey, B. T., Geenen, J. E., Hogan, W. J. (1991). Evaluation of a Voice Recognition System for Generation of Therapeutic ERCP Reports. Gastrointestinal Endoscopy, 37(6):617-620.

McMillan, P. J., Harris, J. G. (1990). Datavoice: A Microcomputer-Based General Purpose Voice-Controlled Data-Collection System. Computers in Biology and Medicine, 20(6):415-419.

Mitchell, J. and Shneiderman, B. (1989). Dynamic versus Static Menus: An Exploratory Comparison, ACM SIGCHI Bulletin, 20(4):33-37.

Newell, A. F. (1992). Wither Speech Systems? Some Characteristics of Spoken Language Which may Effect the Commercial Viability of Speech Technology. In Advances in Speech, Hearing, and Language Processing, JAI Press Ltd., London.

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. and Cohen, P. R. (1991). Discourse Structure and Performance Efficiency in Interactive and Noninteractive Spoken Modalities. Computer Speech and Language, 5(4):297-326.

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.

Poon, A. D. and Fagan, L. M. (1994). PEN-Ivory: The Design and Evaluation of a Pen-Based Computer System for Structured Data Entry, Knowledge Systems Laboratory Technical Report LSK-94-30, Stanford University School of Medicine.

Ramon, T.V. (1995). Information on the NII is not just for Viewing! Digital Equipment Corporation.

Salisbury, M. W., Hendrickson, J. H., Lammers, T. L., Fu, C., and Moody, S. A. (1990). Talk and Draw: Bundling Speech and Graphics. IEEE Computer, 23(8):59-65.

Schmandt, C., Ackerman, M. S., and Hindus, D. (1990). Augmenting a Window System with Speech Input. IEEE Computer 23(8):50-56.

Sears, A., Shneiderman, B. (1991). High Precision Touchscreens: Design Strategies and Comparisons with a Mouse. International Journal of Man-Machine Studies 34:593-613.

Shneiderman, B. (1980). Natural vs. Precise Concise Languages for Human Operation of Computers: Research Issues and Experimental Approaches. In Proceeding of the 18th Annual Meeting of the Association for Computational Linguistics, pp. 139-141, Philadelphia, Pennsylvania.

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

Shneiderman, B. (1992). Designing the User Interface: Strategies for Effective Human-Computer Interactions, page 256, Addison-Wesley, Don Mills Ontario.

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

Smith, N. T., Brian, R. A., Pettus, D. C., Jones, B. R., Quinn, M. L., Sarnat, L. (1990). Recognition Accuracy with a Voice-Recognition System Designed for Anesthesia Record Keeping. Journal of Clinical Monitoring, 6(4):299-306.

U.S. Food and Drug Administration. (1978). Good Laboratory Practice Regulations for Non-Clinical Laboratory Studies. Federal Regulations, 43(247):60015-60019.

Voice Processing Magazine 1994 Buyer's Guide (1993). 5(12):35.

Webber, B. L. (1986). So What can we Talk About Now? In Computational Models of Discourse, MIT Press, Cambridge, Massachusetts, 1983. Reprinted in Readings in Natural Language Processing, Morgan Kaufman Publishers, Inc., Los Altos, California.

Welch, J. R. (1977). Automated Data Entry Analysis. Rome Air Development Center Report, RADC TR-77-306, Griffiss Air Force Base, NY.

Wright, P., Lickorish, A., Milroy, R. (1994). Remembering While Mousing: The Cognitive Costs of Mouse Clicks. SIGCHI Bulletin, 26(1):41-45.

Wulfman, C. E., Isaacs, E. A., Webber, B. L., Fagan, L. M. (1988). Integration Discontinuity: Interface Users and Systems. Tech. Report KSL-88-12, Knowledge Systems Laboratory, Stanford University, Palo Alto, California.

Wulfman, C. E., Rua, M., Lane, C. D., Shortliffe, E. H., Fagan, L. M. (1993). Graphical Access to Medical Expert System: V. Integration with Continuous-Speech Recognition. Methods of Information in Medicine, 32(1):33-46.

Zloof, M. M. (1977). Query-by-Example: A Data Base Language. IBM Systems Journal, 16(4):324-343.