Crossroads
The ACM Student Magazine

Task Integration in Multimodal Speech Recognition Environments

by ***Michael A. Grasso*** *and* ***Tim Finin***

Introduction

The benefits of speech-driven user interfaces have been advocated for several years. 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. Taking this into account, several research efforts have begun to focus on speech as an ancillary input channel in multimodal environments.

One example of this is the effort to combine speech with direct manipulation. Direct manipulation interfaces, made popular by the Apple Macintosh and Microsoft Windows graphical environments, are based on the visual display of objects of interest and the selection by pointing instead of typing **[1]**. For simplicity, the term *speech recognition* will deal with the identification of spoken words, not necessarily natural language recognition, and *direct manipulation* will deal with mouse-driven input. While a mouse-driven interface by itself is not necessarily a direct manipulation interface, there is enough overlap between the two in the context of this discussion.

A complementary model of behavior has been proposed, suggesting that direct manipulation and speech recognition interfaces have reciprocal strengths and weaknesses which could be leveraged in a multimodal user interface. By combining the two modalities, the strengths of one could be used to offset the weaknesses of the other.

> *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* [2].

Put another way, direct manipulation interfaces are believed to be best used for specifying simple actions when all references are visible and references are limited in number. In contrast to this, speech recognition interfaces are thought to be better at specifying more complex actions when references are numerous and not visible. These specific attributes are outlined in the table below.

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

Proposed Applications for Direct Manipulation and Speech

To understand how to leverage an advantage, anecdotal arguments need to be evaluated scientifically. More theoretical work is needed in order to help predict the performance of speech in multimodal environments [3], [4], [5]. The focus of this paper, therefore, is to propose a framework to empirically evaluate the types of tasks that might benefit from a multimodal interface. Before exploring this issue, an overview of speech recognition technology is given. This is followed by theoretical work in task integration. Related work in multimodal speech interfaces is also covered. Finally, a framework for evaluating the types of input tasks that could benefit from multimodal environments is presented.

Speech Recognition Technology

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 [6]. Recognition accuracy was reported to be 98%. Later that decade, a system with similar attributes was developed that recognized consonants and vowels [7]. In the 1960s, research in speech recognition moved to digital computers. This platform provided the basis for speech recognition technology to the present day [8].

Despite rapid progress early on, limitations in computer architectures prevented any significant commercial speech recognition system development. Note that although 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, however, a number of commercial systems have been successfully developed [9]. Despite these advances, true speech processing is still several years away. Therefore, a successful speech-driven system must allow for limitations in the current technology. These limitations include speaker dependence, continuity of speech, and vocabulary size.

Speaker *independent* systems can recognize speech from any speaker. Speaker *dependent* systems must be trained by each individual user, but typically have higher accuracy rates. Speaker *adaptive* systems, a hybrid approach, start with speaker-independent templates and adapt them to specific users over time without explicit training. *Continuous* speech systems can recognize words spoken in a natural rhythm while *isolated word* systems require a deliberate pause between each word. Although more desirable, continuous speech is harder to process, because of the difficulty in detecting word boundaries. Vocabulary size can vary anywhere from 20 words to more than 40,000 words. Large vocabularies cause difficulties in maintaining accuracy, but small vocabularies can impose unwanted restrictions on the naturalness of communication. Often the vocabulary must be constrained by grammar rules which identify how words can be spoken in context. A more thorough review of this subject can be found elsewhere [10].

Along with technical characteristics of speech recognition systems, it is important to understand the human factors of speech as an interface modality. The most significant is that speech is temporary. Once uttered, auditory information is no longer available. This may place extra memory burdens on the user and severely limit the ability to scan, review and cross-reference information. Speech can be used at a distance which makes it ideal for hands-busy and eyes-busy situations. It is omnidirectional and therefore can communicate with multiple users. However, this has implications related to privacy and security. Finally, more than other modalities, there is the possibility of anthropomorphism when using 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 [11].

Task Integration

Input devices like speech and a mouse 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.

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 [12]. The **perceptual structure** is defined as how the input dimensions are perceived by the user. A two-dimensional mouse and a three-dimensional tracker were selected as input devices. Two input tasks with three inputs each were evaluated. In the first task, the inputs were integral (x location, y location, and size) and in the other, the inputs were separable (x location, y location, and color).

Common sense might say that a three-dimensional tracker is a logical superset of a two-dimensional mouse and therefore always as good and sometimes better than a mouse. Instead, the results showed that the tracker performed better when the three inputs were perceptually integral, while the mouse performed better when the three inputs were separable.

The theory of perceptual structures, integral and separable, was originally developed by Garner [13], [14]. The structure has to do with how the dimensions of an input task combine perceptually. The basis for the x location, y location, and size of an object being integral and the x location, y location, and color being separable was taken from this work.

This theory was extended with the hypothesis that the perceptual structure of an input task is key to the performance of multidimensional input devices on multidimensional tasks in a unimodal environment. An appropriate area for additional research is to evaluate the performance of integral and separable multidimensional input tasks in multimodal environments, where two or more modalities are used in concert. The following section includes examples of related work in the area.

Multimodal Multidimensional Task Integration

A number of observations were made by Oviatt and Olsen with respect to how people integrate input from different devices in multimodal environments [15]. Participants were asked to perform data entry tasks using a multimodal speech and handwriting user interface. During the experiment, participants were free to use whichever modality they wanted. It was noted that the most influential factor in predicting the use of integrated multimodal speech and handwriting was contrastive functionality. In other words, participants were most likely to integrate the two modalities in a contrastive way to designate a shift in context or functionality, such as original input versus corrected input, or data versus command.

A project by Cohen used speech and direct manipulation to develop an integrated user interface [16]. Here, the goal was not simply to provide two or more separate modalities with the same functionality, but to integrate them together to produce a more productive interface. For example, along with traditional unimodal operations like "point-and-click," there can be multimodal ones like "point-and-speak." Their intent was to use the strengths of one modality to overcome the weaknesses of the other.

Considering this objective, a prototype multimodal system was developed that used an integrated direct manipulation and natural language interface. Several examples were cited where the combination of language and mouse input was thought to be more productive than either modality alone. For example, natural language allows the use of anaphoric references (pronouns). However, the exact meaning of these references can be ambiguous. When such references were unclear, 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 mouse input 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 point in time. 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** for **the Airborne Warning and Control System (AWACS)** (http://www.boeing.com/dsg.awacs.html) **[17]**. Noting that human communication is multidimensional and that conversations include more than just spoken words, they used a combination of graphics and verbal data where one completes or disambiguates the other. 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.

A similar approach was taken while integrating a natural language interface with a graphical airborne early warning test planning tool at **the Naval Research Laboratory** **[18]**. The use of natural language provided expressive power above and beyond what is possible with direct manipulation. For example, using speech, a user could specify a command, a reference, and destination, such as "Move fighter 14 to station 5." Alternatively, using multimodal input, a user could specify the command and reference only as "Move fighter 14." The destination would then be selected using the mouse to click on a location from a graphical map.

Following intuitive guidelines, these efforts seemed to integrate speech in multimodal, multidimensional input tasks when the input attributes were perceptually separable. Examples of this are when there was a shift in context or function, such as reference identification versus data input, description versus examination in the context of time, and data versus command. This suggests that in an empirical evaluation, performance may improve when perceptually separable attributes are input using different modalities.

Feasibility Study

Preliminary work by the author includes a feasibility study of speech-driven data collection **[19]**. The objective was to determine the feasibility of using speech 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 data using only speech input and computer-generated speech responses. The prototype supported continuous-speech and speaker-dependence with a vocabulary of 900 words based on the Pathology Code Table [20].

After testing the prototype system, the results were evaluated to determine the feasibility of using speech in this application area. The overall accuracy rate for speech recognition was 97%. Additional work is needed to minimize training requirements and improve audible feedback. However, it was concluded that this architecture could be considered a viable alternative for data collection in animal toxicology studies with reasonable recognition accuracy.

Future Work

The prototype software has been modified to support an experimental study of task integration using speech and mouse input. The application domain for the prototype is histopathologic data collection in animal toxicology studies. This type of study is used to evaluate the long-term, low-dose effects of potentially toxic substances, including carcinogens. It is based on a highly structured, specialized, and moderately sized vocabulary that includes several significant hands-busy and eyes-busy restrictions. These and other characteristics make it a typical data collection task, similar those required in biomedical research and clinical trials. The input tasks mainly involve reference identification with little declarative or spatial data entry required, which should minimize any built-in bias toward either modality. Also to remove bias, specific hands-busy and eyes-busy restrictions were removed.

To eliminate up-front training requirements, the new prototype uses the PE500 for speech recognition (**Speech Systems, Inc.** Boulder, CO, USA. Http://www.speechsys.com). It supports speaker-independent, continuous recognition of grammatically constrained vocabularies. The original prototype used either mouse or speech input for data collection, but not both at the same time. The new prototype supports the use of mouse and speech input together. This is to allow for the comparison of multimodal, multidimensional input tasks using speech and the mouse in concert.

An experimental study is under way to evaluate the performance, accuracy, and acceptability of using speech and direct manipulation for various multidimensional input tasks in the context of animal toxicology studies. Around 40 veterinary pathologists, toxicologists, and residents will participate in this study.

Based on the theory of perceptual structures, the literature reports that the performance of multidimensional, unimodal input tasks is effected by whether the inputs are perceived as integral or separable. In addition, users are more likely to switch from one modality to another when there is a change in functionality or context. The objective of this study is to empirically evaluate two questions. The first is whether the speed, accuracy, and acceptance of multidimensional, multimodal input will increase when the attributes of the task are perceived as integral or separable. The second is to determine if integral or separable input tasks using mouse and speech together perform better than mouse-only or speech-only input.

Conclusion

A model of complementary behavior has been proposed based on arguments that direct manipulation and speech recognition interfaces have reciprocal strengths and weaknesses. This suggests that user interface performance and acceptance may increase by adopting a multimodal approach that combines speech and direct manipulation. More theoretical work is needed in order to understand how to leverage this advantage. In this paper, a framework was presented to empirically evaluate the types of tasks that might benefit from such a multimodal interface.

References

1.
Shneiderman, B. ***Sparks of Innovation in Human-Computer Interaction*** , Ablex Publishing Corporation, Norwood, NJ, 1993.

2.
Cohen, P. R., Oviatt, S. L. The Role of Voice in Human-Machine Communication. In *Voice Communication Between Humans and Machines*, pp. 34-75, National Academy Press, 1994.

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

4.
Carbonell, N. Multimodal Human-Computer Interaction. *ACM SIGCHI Bulletin*, 26, 3, (July 1994), pp. 15-18.

5.
Damper, R. I. 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 , page 70, Taylor and Francis, 1993.

6.

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

7.

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

8.

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

9.

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

10.

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

11.

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

12.

Jacob, R. J. K., Sibert, L. E., McFarlane, D. C., Mullen, M. P. Integrality and Separability of Input Devices. *ACM Transactions on Computer-Human Interaction*, 1, 1, (March 1994), pp. 3-26.

13.

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

14.

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

15.

Oviatt, S. and Olsen, E. 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, 1994.

16.

Cohen, P. R. 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, 1992.

17.

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

18.

Marsh, E., Wauchope, K, Gurney, J. O., *Human-Machine Dialogue for Multi-Modal Decision Support Systems*, NCARAI Report AIC-94-032, Navy Center for Applied Research in Artificial Intelligence, Navy Research Laboratory, Washington, DC, 1994.

19.

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

20.

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

About the Authors

Michael Grasso **(grasso@cs.umbc.edu)** is a doctoral student in the Computer Science and Electrical Engineering Department at the University of Maryland, Baltimore County. His research interests include human computer interaction and biomedical computing. He received a B.S. in Microbiology from the University of Maryland, College Park in 1983 and an M.S. degree in Computer Science from the American University in Washington, DC in 1986. He currently works for Segue Corporation, a company he co-founded in 1986 which focuses on biomedical and scientific computing. For more information about him refer to his web site **(http://www.cs.umbc.edu/~mikeg)**.

Tim Finin (finin@cs.umbc.edu, **http://www.cs.umbc.edu/~finin**) is a Professor of Computer Science and Electrical Engineering at the University of Maryland Baltimore County. He has had over 25 years of experience in the applications of Artificial Intelligence to problems in database and knowledge base systems, intelligent information systems, expert systems, natural language processing, intelligent interfaces and robotics. Prior to joining the University of Maryland, he was a Technical Director at the Unisys Center for Advanced Information Technology, a member of the faculty of the University of Pennsylvania, and on research staff of the MIT AI Lab. He holds an SB degree in EE from MIT and a PhD in Computer Science from the University of Illinois.