An Empirical Study of Dual-Modal Information Presentation

Xiaowen Fang
DePaul University
xfang@cti.depaul.edu

Jacek Brzezinski
DePaul University
JBrzezinski@cti.depaul.edu

Kimberly Watson
DePaul University
kwatson@cti.depaul.edu

Shuang Xu
DePaul University
sxu@cti.depaul.edu

Susy Chan
DePaul University
schan@cti.depaul.edu

ABSTRACT

Dual-modal interfaces with both visual and auditory output are becoming important, especially for applications using small-screen displays and for user access under mobile conditions. Our research investigated the effectiveness and feasibility of a dual-modal information presentation, "Visual + Auditory". Based on the multiple-resource human attention model (Wickens, 1980, 1984), this study hypothesizes that additional auditory information presented during a web browsing process can be perceived by users and will not negatively impact users' performance on the browsing task. A controlled experiment was conducted to test the hypothesis. Display mode was the independent variable, and the two treatments were regular visual display and visual plus auditory. The dependent variables were user satisfaction and user task performance. The hypothesis was fully supported by the experiment. The findings suggest that users can perceive auditory information while visually browsing textual information.

Keywords

Dual-modal interface, visual, auditory, information presentation

INTRODUCTION

The convergence of mobile Internet and wireless communication technology has promised users anytime, anywhere access to information for work and personal communication. However, many technology constraints hinder such access through a mobile device. Two important constraints are the device's small screen size and its mobile use. Compared to desktop or laptop computers, mobile devices typically have a very small screen on which only a very limited amount of information can be presented. When the device is used on the move, it makes reading textual information much more difficult (Chan, Fang, & Brzezinski et al., 2002).

Multimodal interfaces are considered by many to have a very bright future. It has been predicted that in the long run more businesses will use voice applications to supplement or complement text-based e-commerce applications (Economist Technology Quarterly, 2002). Several factors drive this trend: (1) A voice-based interface may increase the potential customer base because there are still more telephones than Web-enabled devices; (2) The screen size is limited on wireless devices; and (3) Customers use their telephones more frequently because they prefer to communicate using natural language. Previous research has examined the effects of presenting information in both auditory and visual modes. In general, the major advantage of using speech in an interface is its universality. The main disadvantage is that human beings can only process voice output at a relatively slow speed (Streeter, 1988). Auditory presentation should be used if the task is performed in continuous motion (Proctor & Van Zandt, 1994). Based on Wickens' (1980, 1984) multiple-resource human attention model, two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. In other words, humans could accept information without interference from two completely different

channels: visual and auditory, at the same time. Therefore, if voice output is integrated with visual presentation on a mobile device, problems related to limited screen space and mobility could be addressed by a multimodal interface. Modality integration will require the split of retrieved and summarized information into visual and auditory channels.

The objective of this study is to investigate the effectiveness and feasibility of a dual-modal information presentation, "Visual + Auditory," which outputs information through the auditory channel in addition to the regular visual display. Findings from this research could contribute to the establishment of guidelines for the design of dual-modal interfaces.

BACKGROUND LITERATURE

Human Attention

The topic of attention has long been of interest to researchers. Proctor and Van Zandt (1994) distinguish human attention in three aspects: selective attention that concerns human ability to focus on certain sources of information and ignore others; divided attention that involves human ability to divide attention among multiple tasks; and the amount of mental effort required to perform a task. Several models of attention have been proposed. Bottleneck models specify a particular stage in the information-processing sequence at which the amount of information that humans can attend to is limited. In contrast, resource models view attention as a limited-capacity resource that can be allocated to one or more tasks, rather than as a fixed bottleneck. Among various attention models, multiple-resource models propose that there is no single attention resource. Rather, several distinct subsystems each have their own limited pool of resources. Wickens (1980, 1984) proposes a three-dimensional system of resources consisting of distinct stages of processing (encoding, central processing, and responding), codes (verbal and spatial), and input (visual and auditory), plus output (manual and vocal) modalities. The model assumes that two tasks can be performed together more efficiently to the extent that they require separate pools of resources.

Human Working Memory

Figure 1 shows a diagram of the working memory model proposed by Baddeley (1986). In this model, acoustic or phonological coding is represented by the phonological loop, which plays a role in vocabulary acquisition, learning to read, and language comprehension. The model also includes visual coding, in the form of the visuo-spatial "sketch pad." This sketch pad is assumed to be responsible for visual imagery. The central executive is an attentional control system that supervises and coordinates the visuo-spatial and phonological subsystems. According to Baddeley's working memory model, visual imagery information and acoustic verbal information can be simultaneously held in separate storage systems, which can be further integrated by the central executive with minimum cognitive cost. Therefore, tasks should not interfere with each other if they use different subsystems. Many studies have reported evidence supporting this model. Mousavi, Low, and Sweller (1995) have reported findings on the use of a partially auditory and partially visual mode of presentation for geometry examples. The effects of presentation modality suggest that working memory has partially independent processors for handling visual and auditory materials. Effective working memory may be increased by presenting materials in a mixed rather than a unitary mode.

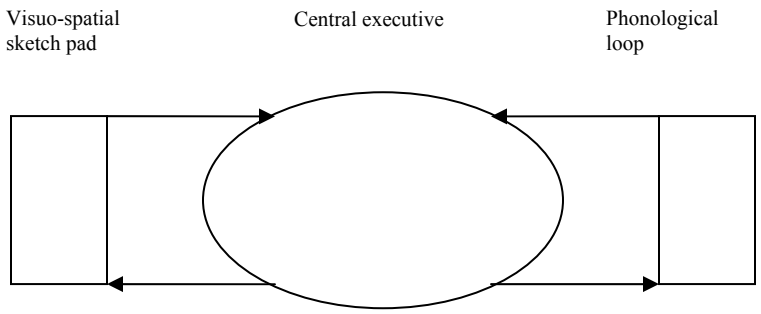

Figure 1. Working memory model proposed by Baddeley

Visual and Auditory Interfaces

The majority of displays encountered in human-machine systems are either visual or auditory. In general, because spatial discriminations can be made most accurately with vision, spatial information is best conveyed through visual displays (Proctor & Van Zandt, 1994). Likewise, auditory displays work best with temporal information because temporal organization is a primary attribute of auditory perception.

Several studies have investigated the effects of using text, voice, or text combined with voice output modes. In comparing speech with text, Streeter (1988) points out that the major advantage of using speech in an interface is its universality, because everyone understands spoken language and users can be mobile. One notable disadvantage is that voice delivers information at less than half the rate that text can be read. In a study of the effects of media (pictures, audio, and print) on student learning, Nugent (1982) found that student learning could be improved by incorporating additional channels such as pictures and audio when the same content was presented in three channels. When different information was presented in the visual and audio modes, however, student learning was not improved by the addition of new channels but the presence of visual did not interfere with processing the audio and vice versa. A similar study conducted by Baggett and Ehrenfeucht (1983) suggests that there is no competition for resources when related information is presented simultaneously in visual and auditory channels.

Sipior and Garrity (1992) indicate that presentation with a mix of audio and visual accompaniments improve receptiveness attributes such as perception, attention, comprehension, and retention. DeHaemer and Wallace (1992) suggest that the visual and audio modes of receiving information appear to be non-interfering and may enhance performance for certain tasks. They observed the effect of voice output on computer-supported decision making, where voice instructions were used to solve a visual decision problem, and found an interaction effect between user decision style and the use of computer synthetic voice. In comparing voice and text annotation in co-authored documents in terms of interactivity and expressiveness, Chalfonte, Fish, and Kraut (1991) found that voice was preferred for addressing higher level issues in suggesting document modifications, but text was preferred for more detailed and lower level comments. Archer, Head, Wollersheim, and Yuan (1996) designed an interface to study user preferences and the effectiveness of output modes. Their findings show that adding text to voice output improves the perceived acceptability of voice, but adding voice to text does not alter the perceived acceptability of text. When the same information was presented in both modes, text mode was most efficient in performing information search, followed by voice mode, and text plus voice mode.

Speech output has also been used widely to assist sight-impaired users. Schlosser, Belfiore, and Nigam (1995) point out that the presentation of additional auditory stimuli in the form of synthetic speech is effective in assisting individuals with mental retardation to learn associations of graphic symbols with spoken words. Gorenflo and Gorenflo (1994) show that attitudes toward the augmented communicator are more favorable in terms of evaluation and potential interaction when the synthetic voice is easier to listen to. Krell and Cubranic (1996) developed a special browser "V-Lynx" with voice output for sight-impaired users. The "V-Lynx" browser reads the first sentence in a paragraph for quick scanning of the document, conveys the document structure (headings, emphasized text, lists, and hyperlinks), and allows for easy navigation while inside and between documents.

Oviatt and Olsen (1994) have studied how people use computers when allowed to use multimodal interfaces for data entry, retrieval, and navigation. The most important factor in predicting the use of multimodal interface is contrastive functionality. For example, users utilize one modality for data input but another for input corrections. A project conducted by Cohen (1992) involved the integration of keyboard, screen, pointing device, and auditory channel. Completion of a task required the utilization of at least two channels such as auditory and keyboard.

Media Richness Theory (MRT), also referred to as Information Richness Theory, has been proposed to explain the impact of computer technology on communication (Daft & Lengel, 1986). Richness refers to the capacity of the medium to process information. Communication across various media differs based on the bandwidth or number of cue systems available within the media. The theory holds that there are four factors that determine the richness of the media: 1) ability of the medium to transmit multiple cues, such as body language and voice inflection; 2) availability of immediate feedback; 3) facility to use natural language; and 4) capability of infusing the message with personal feelings and emotions. Rich media has a higher capacity to facilitate shared meaning than lean media. Richness (or leanness) is a property of the technology that serves as the communication media. Because the richness property is intrinsic to the technology, media and messages should be matched. Simple, unambiguous messages are suitable for lean communication channels, such as email, while more emotional and ambiguous messages are best conveyed via rich channels, such as face-to face communication. Chidambaram and Jones (1993) have confirmed in their study that face-to-face meetings, compared to other types of media, have a broader bandwidth, permit the exchange of richer information, and offer a more natural setting for group communication. The multimodal interfaces to be investigated in this study will certainly add richer information to the normal textual computer display through auditory channel and thus might be able to convey more complicated information.

PROPOSED DUAL-MODAL INFORMATION PRESENTATION

Based on the prior research findings, a dual-modal information presentation, "Visual + Auditory Information" was proposed. In this presentation format, a regular document is displayed in normal visual mode while additional information is presented as voice output. The multiple-resource human attention model proposed by Wickens (1980 & 1984) suggests that two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. Based on this model, users might be able to allocate resources to attend the auditory information while browsing a textual document. Baddeley's working memory model (1986) indicates that visual imagery information and acoustic verbal information can be simultaneously held in separate storage systems, which can be further integrated by the central executive with minimum cognitive cost. When users receive brief auditory information during a browsing process, information from the two different modalities (visual and auditory) might be stored in different subsystems based on Baddeley's working memory model. If the time spent on processing the auditory information is short enough, users might be able to temporarily remember the visual context of the browsing task in working memory and then smoothly resume the browsing without too much disruption after the voice information is completed. Therefore, users may be able to perceive brief information from the auditory channel while they are retrieving information visually. The following hypothesis will be used to test the effectiveness of this dual-modal information presentation.

Hypothesis: Additional auditory information presented during a web browsing process can be perceived by users and will not negatively impact users' performance on browsing tasks.

A simple T-test comparing "Regular Visual Display" (text only) and "Visual + Auditory Cues" (text/voice) will be used to test this hypothesis. If there is no significant difference in the average number of correctly answered text-based questions between the "Regular Visual Display" and the "Visual + Auditory Cues" groups, then additional auditory information does not negatively impact users' performance on browsing tasks. If the average number of correctly answered questions that are related to the voice cues heard by the "Visual + Auditory Cues" group is significantly greater than zero, then additional auditory information can be perceived by users during a web browsing process.

METHOD

A web site containing generic curriculum information was developed for this experiment. Additional curriculum information was designed and pre-recorded for auditory presentation. The user's task was to browse the web site and listen to the auditory presentation (if any), find information relevant to pre-defined task questions based upon both the text and auditory output, and answer the task questions. Participants performed tasks on a personal computer (PC). A set of questions about the web site and the additional information delivered in auditory mode were designed to measure the user's performance on information retrieval tasks.

The only independent variable was the information presentation mode. There were two modes: "Regular Visual Display" and "Visual + Auditory Cues". In the "Regular Visual Display" mode, all information was presented visually with no auditory cue. The task for participants was to browse the textual information contained in the Web site and answer questions based on the texts. In the "Visual + Auditory Cues" mode, additional generic curriculum information was randomly presented through auditory channel while participants were browsing the textual content of the Web site. In addition to the primary browsing tasks, participants in this group also needed to listen to the auditory output that might be presented at any time during the experiment and answer questions derived from auditory information. Figure 2 shows a screen shot of the "Visual + Auditory Cues" mode.

Dependent variables included the number of correctly answered questions related to the text-based web site, the number of correctly answered questions related to the voice cues, and satisfaction. User performance was measured by the number of correctly answered questions related to the text-based web site and the number of correctly answered questions related to the voice cues (if applicable). Satisfaction was measured by a questionnaire based on original scale items developed by Davis (1989).

As of April 8, 2004, twenty six (26) participants were recruited from a university in the US Midwest region, which hosts a variety of different age groups, ethnicities, computer experience levels, and knowledge backgrounds. Participants were randomly assigned to the two groups: "Regular Visual Display" and "Visual + Auditory Cues". Out of the twenty six

participants, sixteen were females and fourteen were native English-speakers. Both female participants and native English-speakers were evenly distributed in the two experiment groups.

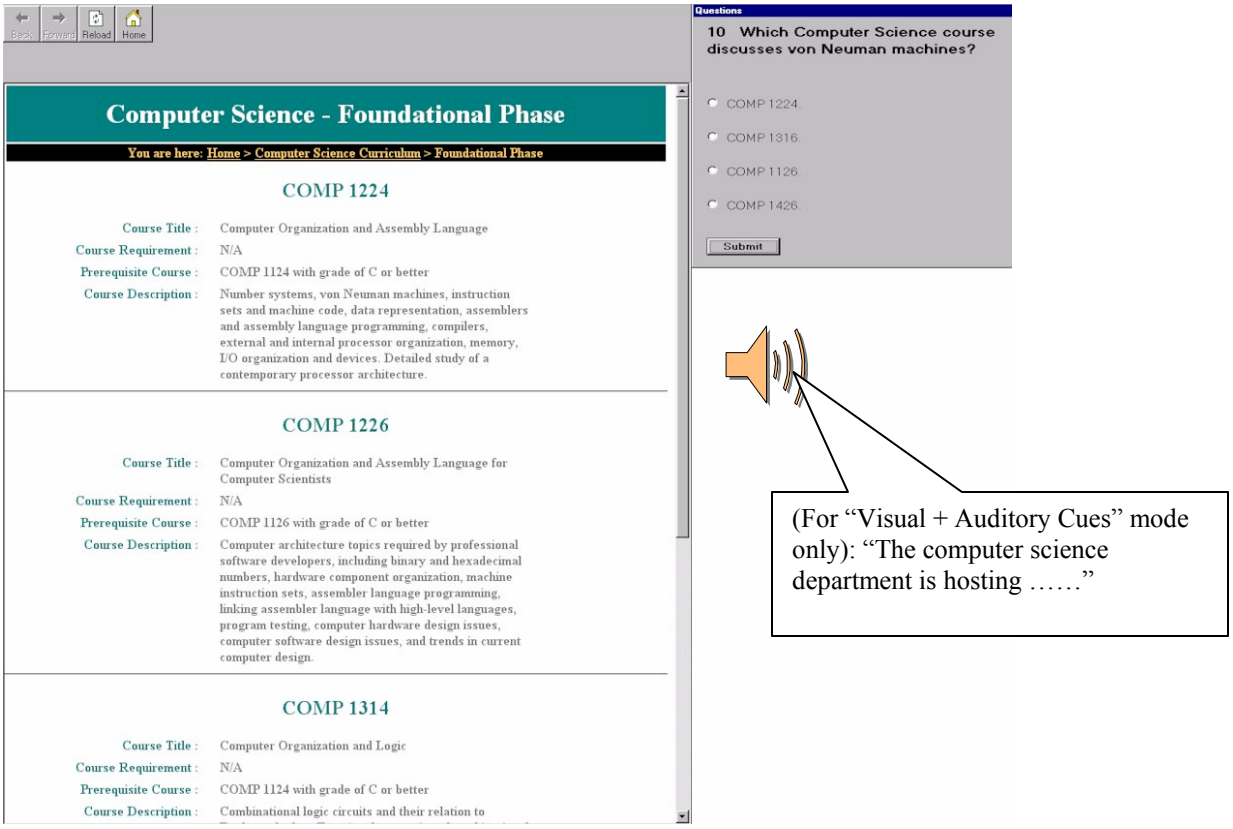

Figure 2. Screen Shot of "Visual + Auditory Cues" Presentation Mode

Each participant was asked to sign a consent form before participating in the experiment. Upon completion of a pre-experiment questionnaire, each participant received instructions for using the experiment browser and performing the required tasks. These instructions were presented via computer and in hard copy format. The participant then started a training session in which he or she could browse a sample web site with visual (and auditory) information similar to the version the user would encounter in the experiment. Upon successful completion of the training test, the participant began the experiment tasks. Considering that fatigue factor might affect participants' understanding capability and listening comprehension, the total duration of the experiment was limited to 30 minutes. All participants were asked to correctly answer as many questions as they could and as quickly as possible in the given period. A large number of questions have been developed so that no subject could finish all of them. Subjects were informed of the time limit prior to starting the experiment in order to focus their attention on the tasks at hand, and the browsing task was programmed to automatically terminate at the 30 minute mark. Following the completion of the experiment tasks or the expiration of the experiment time allowed, the participant was asked to complete a survey about his or her perceptions of the ease of use (Davis, 1989), usefulness (Davis, 1989), and the perceived control they enjoyed over the information display mechanism. Participants in the "Visual + Auditory Cues" group were then debriefed to provide additional information about their experience in processing and using the voice and text information. The debriefing sessions were recorded for data analysis.

The data collection is still in progress. Preliminary results are reported in the following section.

RESULTS AND DISCUSSION

The intention of the hypothesis was to test the effectiveness and feasibility of a dual modal information presentation. It postulates that additional auditory information presented during a web browsing process can be perceived by users and will not negatively impact users' performance on browsing tasks. The dependent variables were the number of correctly answered questions related to the text-based web site, the number of correctly answered questions related to the voice cues, and satisfaction. The mean values, standard deviations, and t-test results of each dependent variable for both experiment groups are presented in Table 1. The satisfaction score was the sum of scores given to individual questions. Besides the three dependent variables, the accuracy was also analyzed. It was defined as the division of number of correctly answered questions over the total number of questions that the participant answered.

Variables	Regular Visual Display (n = 13)		Visual + Auditory Cues (n = 13)		*t*	Pr > |*t*|
	Mean	Std.	Mean	Std.		
Number of correctly answered questions related to the text-based web site	21.3	10.10	22.2	7.77	-0.24	0.81
Accuracy for text-based questions	0.744	0.1187	0.785	0.0927	-0.99	0.33
Satisfaction	50.2	14.34	52.7	8.70	-0.55	0.59
Number of correctly answered questions related to the voice cues	N/A	N/A	10.2	3.02	12.11	0.0001
Accuracy for voice-based questions	N/A	N/A	0.963	0.0596	58.27	0.0001

Table 1. Comparison between "Regular Visual Display" and "Visual + Auditory Cues" Presentations

As shown in Table 1, no significant differences were found between the two experiment groups in number of correctly answered questions related to the text-based web site (*t(24) = -0.24, p = 0.81*), accuracy (*t(24) = -0.99, p = 0.33*), and satisfaction (*t(24) = -0.55, p = 0.59*) at $\alpha = 0.05$ level. This result suggests that presenting additional information through auditory channel didn't significantly degrade participant's performance in the Web browsing task and didn't significantly reduce participant's satisfaction.

Another t-test indicates that the mean of number of correctly answered questions related to the voice cues in the "Visual + Auditory Cues" group was significantly greater than zero (*t(12) = 12.11, p = 0.0001*). The mean of accuracy for voice-based questions was also significantly greater than zero (*mean = 0.963, t(12) = 58.27, p = 0.0001*). This is a clear sign that participants in the "Visual + Auditory Cues" group did successfully receive some information delivered through the auditory channel.

Therefore, the hypothesis was fully supported by this experiment. The results agree with prior research findings. Based on the multiple-resource human attention model proposed by Wickens (1980 & 1984), two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. According to Baddeley's working memory model (1986), tasks using different subsystems (visuo-spatial sketch pad and phonologic loop) in the working memory should not interfere. While users are visually browsing information, they may be able to receive brief information from the auditory channel.

The short debriefing interviews conducted for the "Visual + Auditory Cues" group at the end of the experiment reflect the following patterns:

- Participants could recognize and remember the key words and phrases contained in the voice messages.
- The first few words in the voice messages were distracting but caught the participant's attention.

- There seemed to be a filtering process when a voice message started. During this filtering process, a participant would decide whether or not to listen to the voice information. If the voice information appeared to be relevant to the primary browsing task, the participant would more likely pay attention to it.
- Participants felt that information irrelevant to the primary browsing task would cause more distraction than relevant information and thus might hinder the browsing task.

CONCLUSIONS

In this study, a dual modal information presentation was proposed and tested through a controlled experiment. The preliminary findings from this study suggest that users can receive information presented through auditory channel while they are visually browsing textual information. Hence, it is possible to use multimodalities (visual plus auditory modes) to present more information than a single modality could. These findings could be applied in the multimodal interface design of mobile applications. However, this study represents the first step towards this direction and the sample size is still small at this point. More participants will be recruited. Future study must address what information should be presented in the visual mode and in the auditory mode.

REFERENCES

1. Archer, N., Head, M., Wollersheim, J., & Yuan, Y. (1996). Investigation of voice and text output modes with abstraction in a computer interface. *Interacting with Computers*, *8*(4), 323-345.
2. Baddeley, A. (1986). Working memory. New York: Oxford University Press.
3. Baggett, P. & Ehrenfeucht, A. (1983). Encoding and retaining information in the visuals and verbals of an educational movie. *Educational Communication and Technology Journal, 31* (1), 23-32.
4. Chidambaram, L. & Jones, B. (1993). Impact of communication medium and computer support on group perceptions and performance: A comparison of face-to-face and dispersed meetings. MIS Quarterly, 17(4), 465-491
5. Chalfonte, B., Fish, R., & Kraut, R. (1991). Expressive richness: a comparison of speech and text as media for revision. In *Proceedings of CHI'91*, 21-26, Addison Wesley.
6. Chan, S., Fang, X., Brzezinski, J., Zhou, Y., Xu, S., & Lam, L. (2002). Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research, 3*(3), 187-199.
7. 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*, 143-149, ACM Press.
8. Daft, R.L. & Lengel, R.H. (1986). Organizational Information Requirements, Media Richness and Structural Design. Management Science, 32(5), 554-571.
9. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.
10. DeHaemer, M. & Wallace, W. (1992). The effects on decision task performance of computer synthetic voice output. *International Journal of Man-Machine Studies*, *36*, 65-80.
11. Economist Technology Quarterly (2002). The power of voice. *Economist Technology Quarterly*, December, 25-26.
12. Gorenflo, C. & Gorenflo, D. (1994). Effects of synthetic voice output on attitudes toward the augmented communicator. *Journal of Speech & Hearing Research, 37*(1), 64-68.
13. Krell, M. & Cubranic, D. (1996). V-Lynx: bringing the World Wide Web to sight impaired users. In *Proceedings of ASSETS'96,* 23-26, New York, NY: ACM.
14. Mousavi, S. Y., Low R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. Journal of Educational Psychology, 87 (2), 319 – 334.
15. Nugent, G. (1982). Pictures, audio, and print: symbolic representation and effect on learning. *Educational Communication and Technology, 30*(3), 163-174.
16. Oviatt, S. & Olsen, E. (1994). Integration themes in multimodal human-computer interaction. In *Proceeding of the International Conference on Spoken Language Processing, 2,* 551-554, Acoustical Society of Japan.
17. Proctor, R. & Van Zandt, T. (1994). *Human factors in simple and complex systems*. Needham Heights, MA: Allyn and Bacon.

18. Schlosser, R., Belfiore, P., & Nigam, R. (1995). The effects of speech output technology in the learning of graphic symbols. *Journal of Applied Behavior Analysis, 28*, 537-549.
19. Sipior, J. & Garrity, E. (1992). Merging expert systems with multimedia technology. *Data Base*, *23*(1), 45-49.
20. Streeter, L. (1988). Applying speech synthesis to user interfaces. In Helander, M. (ed.) *Handbook of Human-Computer Interaction,* 321-343, New York, NY: Elsevier Science Pub.
21. Wickens, C. (1980). The structure of attentional resource. In R. S. Nickerson (ed.), *Attention and Performance VIII,* 239-257, Hillsdale, NJ: Lawrence Erlbaum.
22. Wickens, C. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (eds), *Varieties of Attention*, 63-102, New York, NY: Academic Press.