Dual-modal Presentation of Hierarchical Relationship in Texts

Shuang Xu, Xiaowen Fang, Jacek Brzezinski,and Susy Chan

School of Computer Science, Telecommunications, and Information Systems
DePaul University, 243 South Wabash Avenue, Chicago, IL 60604, USA
{sxu, xfang, jbrzezinski, schan}@cti.depaul.edu

Abstract

This study focuses on how to design a visual-auditory information presentation to: (1) minimize the interference in information processing between the visual and auditory channels; and (2) improve the effectiveness of mental integration of information from different modalities. Baddeley's working memory model suggests that imagery spatial information and verbal information can be concurrently held in different subsystems within human working memory. Based on this model and research on human attention, this study proposes a method to convert textual information of hierarchical relationship into a “graphics + voice” representation and hypothesizes that this dual-modal presentation will result in superior comprehension performance and higher satisfaction as compared to pure textual display. Simple T-tests will be used to test the hypothesis. Results of this study will benefit interface design of generic computer systems by alleviating information overloading in the visual display. Findings may also help to address usability problems associated with small-screen computers.

1 Introduction

Complex visual working environments, overwhelming information output, and mobile information access on small-screen handheld devices lead interface designers to use the auditory channel as a supplementary means of information delivery. The benefit of delivering information across different sensory modalities is often justified by the independent nature of multi-modal information processing, which assumes that there is no interference between tasks and thus no degradation in performance (Cook, Crammer, Finan, Sapeluk & Milton, 1997). However, research in cognitive psychology shows that visual and auditory perceptual processing is closely linked (Eimer, 1999). Problems related to memory and cognitive workload are found in current applications with voice-based interface (Cook et al., 1997). For instance, mental integration of disparate information from different modality channels causes a heavy cognitive memory load. As transient auditory information, speech presentation may impose a greater memory burden. Also, switching attention between modalities may be slow and have a high cost.

The objective of this research is to develop a dual-modal interface that: (1) minimizes the interference in information processing between visual and auditory channels; and (2) improves the effectiveness of mental integration of information from different modalities. This study focuses on the dual-modal presentation of textual information that describes hierarchical or tree structures. Results of this research will facilitate the human-computer interaction via a visual-auditory information presentation, which does not cause extra cognitive workload. Design guidelines could be generated from the findings of this study for a more effective dual-modal information presentation on generic computer systems.

2 Literature review

To develop an effective dual-modal information presentation, we have examined prior research in human attention, working memory, visual and auditory interfaces, and graphical representation of texts.

2.1 Human attention

The interference encountered during multi-modal information perception stems from the allocation of limited attentional resources to concurrent sensory information processing. Researchers have proposed several theoretical attention models to explain the mechanism of resource allocation: bottleneck models, resource pool models, and multiple resource pool models. Bottleneck models (Broadbent, 1958) specify that only a limited amount of information can be brought from the sensory register to working memory. From a different perspective, single resource pool models (Kahneman, 1973) view attention as a resource with limited capacity, which can be allocated to perform information processing tasks. This limited resource pool theory explains that when a certain task demands too much attention, performance on other concurrent tasks will be noticeably degraded. The multiple-resource models assume that instead of sharing a single common attentional resource pool, several distinct subsystems can have their own limited resource pools (Navon & Gopher, 1979; Wickens, 1980 & 1984). Wickens describes a three-dimensional system of resource utilization that consists of distinct processing stages (encoding, central processing, and responding), codes (verbal and spatial), and input (visual and auditory) and output (manual and vocal) modalities. According to this model, two or more tasks can be performed together efficiently as long as they require separate pools of resources.

Allocation of attentional resources during complicated time-sharing tasks across multiple modality channels has long been of interest to cognitive psychology researchers. Research shows that introducing auditory channel into prototypes of civil and military cockpits has resulted in degraded performance (Cook et al., 1997). One explanation is that the total amount of attentional resources is limited. When demanded simultaneously by multi-modal information processing tasks, resources allocated to non-dominant channel decrease, as compared to single-modal information processing. Another explanation is that mental integration of different multi-modal information causes a heavy cognitive load in working memory. If this integration is critical to understanding information received from different sensory channels, performance will degrade.

Cook et al (1997) suggest that speech-based interfaces could be used in a restricted, well-defined task to manipulate the demand on central resources by changing the nature of visual discrimination task and the demand on memory. Wickens and Ververs (1998) examined the effects of display location and image intensity on flight path performance. Their findings suggest that attention is modulated by tasks, which are consistent with the limited attentional resources assumption. Faletti and Wellens (1979) explored the seemingly uneven weighing systems for concurrent information processing across different modalities. They believe that approach-avoidance tendencies in response to specific combinations of design elements might be predicted by developing a formula to integrate environmental information. The use of cell phones in automobiles has increased the public concerns for safety issues. Studies on voice-based car-driver interfaces indicate that performing other tasks while driving takes away from a driver's limited attentional resources. An effective multi-modal interface used in automobiles should minimize the driver's investment in attention, and minimize interference and distraction (Starner, 2002; Siewiorek, Smailagic & Hornyak, 2002; Cellario, 2001; Guglielmetti, 2003; Titsworth, 2002).

The above research findings indicate that both the allocation of attentional resources and interactions between information perceived via visual and auditory channels significantly affect a user's comprehension of a dual-modal interface.

2.2 Working memory

Baddeley (1986) has proposed a working memory model that depicts three components: central executive, visuo-spatial sketchpad, and phonological loop (see Figure 1)

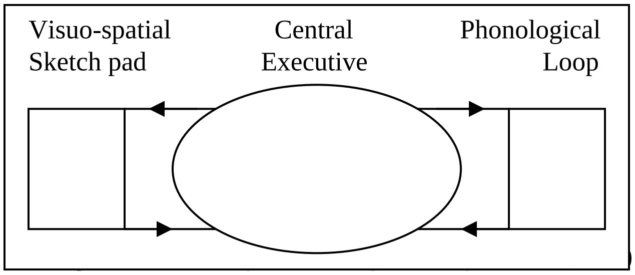

Figure 1. Baddeley's working memory model

According to this model, human working memory contains two subsystems for storage: phonological loop and visuo-spatial sketchpad. Acoustic or phonological coding is represented by the phonological loop, which plays an important role in reading, vocabulary acquisition, and language comprehension. The visuo-spatial sketchpad is responsible for visual coding and handling spatial imagery information in analog forms. The phonological loop and visuo-spatial sketchpad are able to simultaneously hold verbal and imagery information without interference. Central executive is the control system that supervises and coordinates the information retrieved from the two storage subsystems for further integration.

Baddeley's working memory model has been confirmed by a considerable number of experiments. Mousavi, Low, and Sweller (1995) show that students' performance is significantly improved when the verbal representation and image representation of a geometry problem are respectively presented in auditory and visual mode. These researchers further suggest that, because the phonological loop and visuo-spatial sketchpad can hold distinct information independently, working memory might be effectively increased by distributing information in visual and auditory modalities. An earlier research study by Woodhead and Baddeley (1981) indicates that people who are good at recognizing faces also show better performance in recognizing paintings, but do not differ in recognizing words. Their finding implies image memory is separate from verbal memory.

2.3 Visual and auditory information presentation

Introducing voice into the traditional visual interface presents new challenges. The different nature of visual and auditory signals, as well as human's perception processes toward different sensory information, can affect the effectiveness of users' comprehension of integrated multi-modal information. Research in two areas is particularly relevant: (a) comparison between visual and auditory information presentation; and (b) multi-modal interfaces that affect people's comprehension and learning effectiveness.

When comparing visual and auditory information representation, prior research shows that voice is more informal and interactive for handling the complex, equivocal and emotional aspects of collaborative tasks (Chalfonte, Fish & Kraut, 1991). As Streeter (1998) indicates, universality and mobile accessibility are major advantages of speech-based interface, whereas its disadvantage is the slow delivery rate of voice information. Archer, Head, Wollersheim, and Yuan (1996) compared the user's preferences and the effectiveness of information delivery in visual, auditory, and visual-auditory modes. They suggest that information should be organized according to its perceived importance to the user, who should also have flexible information access at different levels of abstraction.

Multi-modal interfaces have been widely used as a support of collaborative work, as well as in teaching systems. Researchers (Nardi et al., 1993) indicate that the integration of video information and other data sources (e.g., aural input, time-based physical data, etc.) help surgeons choose the correct action and interpretation during remote medical operations. Research on interaction between sound, written words, and the image of objects shows that when different sources of information are integrated, a learner's cognitive overload remains light and does not limit learning (Dubois & Vial, 2000). Stock, Strapparava, and Zancanaro (1997) show that hypertext and digital video sequences help users explore information more effectively. By exploring the integration of captioning, video description, and other access tools for interactive learning, Treviranus and Coombs (2000) demonstrated how to make the learning environment more flexible and engaging for students. Dubois and Vial (2000) suggest that several factors affect the effectiveness of integration of multi-modal information. These factors include not only the presentation mode, the construction of co-references that interrelate to the different components of the learning materials, but also the characteristics of the task.

2.4 Graphical representation of texts

To design an effective dual-modal information presentation based on Baddeley's working memory model, it is important to understand how textual information should be converted to imagery/graphical and verbal representations.

Researchers in cognitive psychology have been interested in the knowledge representation via mental images for decades. Studies (Minsky, 1975; Kosslyn & Shwartz, 1977) suggest that mental imagery information is generated from a hierarchically presented structure from long-term memory. This hierarchical knowledge representation includes the skeletal shape of the object and details or other components that are attached to the skeleton. The association between the new information and the components in this tree structure significantly affects the effectiveness of people's comprehension. Other researchers indicate that it is useful to be able to see the entire hierarchy while focusing on a particular part so that the relationship of parts-whole can be seen and the focus can be guided to other parts in a smooth and continuous way (Lamping, Rao & Piroli, 1995). For example, Cone Tree display in lucid forms complicated data hierarchies that might be otherwise invisible to the user (Robertson, Mackinlay & Card, 1991).

Schema (or script, frame) has also been widely used in knowledge representation (Proctor & Van Zandt, 1994; Johnson-Laird, 1983, 1989). Schemas are frameworks that depict conceptual entities, such as objects, situations, events, actions, and the sequences between them. Schemas not only represent the structure of a person's interest and knowledge, but also enable a person to develop the expectancy about what will occur. Thus schematic theory (Anderson & Pearson, 1984) predicts that content familiarity should enhance comprehension by providing an abstract knowledge framework for incoming information. On the other hand, dual coding theory (Paivio, 1986 & 1971) predicts that concrete language should be better comprehended and easily integrated in memory than abstract language.

Mayer's empirical studies (1989 & 1990) indicate that an effective illustration model should use images or diagrams to reorganize and integrate the acquired information. The illustration must be able to guide user's selective attention towards the key items in the presented information. These key items include not only the major entities (such as objects, states, actions, etc.), but also the relationship among them. Mayer (1991) further indicates that an explanative illustration can be most effective when the (visual) animation and (auditory) narration are presented concurrently.

Based on the above discussions, we propose a dual-modal information presentation that presents the hierarchical relationship depicted in texts as tree-structured diagrams, and presents the remaining textual information as voice message. The following section discusses this dual-modal presentation in greater details.

3 Proposed dual-modal information presentation

Based on Baddeley's working memory model, it is assumed that the effectiveness of human information processing can be improved if the verbal representation and the imagery/graphical representation of certain textual information are presented via auditory and visual output, respectively. As shown in Figure 2, if the verbal presentation of the original textual information is presented via auditory channel, the verbal information will be temporarily stored in the auditory sensory register, then sent to and processed in the phonological loop in working memory. Meanwhile, information perceived from the graphical presentation will be stored in the visual sensory register and then transferred to visuo-spatial sketchpad. Verbal and graphical information that are concurrently stored in working memory could be respectively retrieved from the phonological loop and visuo-spatial sketchpad, and then integrated by the central executive for comprehension.

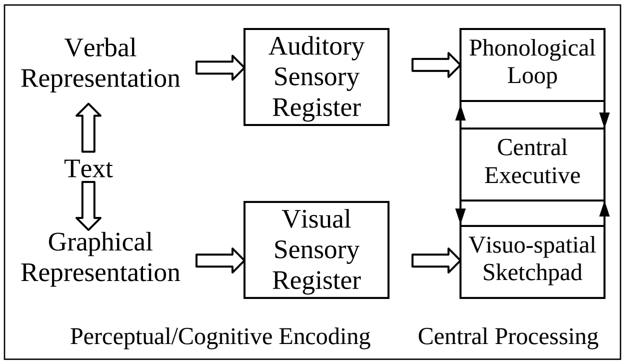

Figure 2. Splitting Textual Information

In this proposed dual-modal presentation (see Figure 3), hierarchical relationship contained in texts will be extracted and presented as a tree structure, with related entities indicated as nodes of this tree. The remaining textual information will be delivered through the auditory channel. The following hypothesis is proposed to test the effectiveness of this dual-modal information presentation.

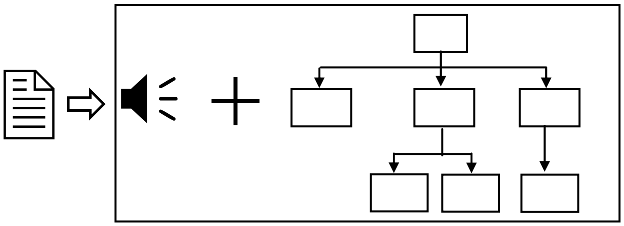

Figure 3. Proposed Dual-modal Presentation

Hypothesis: The dual-modal presentation of hierarchical relationships will improve user's comprehension of information and result in higher user satisfaction as compared to pure textual display.

According to Baddeley's model, pure visual display of textual information will be processed entirely in the phonological loop. Non-speech verbal input must go through a sub-vocal rehearsal to be converted to speech input and temporarily saved in the phonological loop of working memory before further processing (see Figure 4).

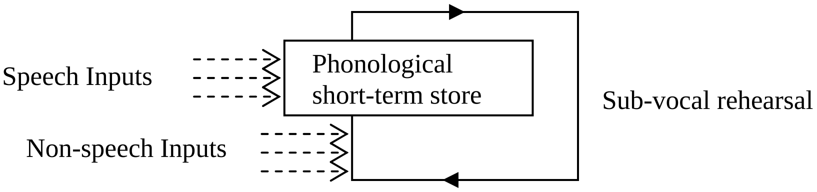

Figure 4. Structure of Phonological Loop

In the proposed dual-modal presentation, the graphical information might be perceived and held in the visuo-spatial sketchpad while the speech input is perceived and directly stored in phonological loop. Therefore, by concurrently utilizing the two subsystems in working memory to process the same amount of information, a reduced cognitive workload is expected during information processing. Research in human attention has shown that many voice-based interfaces caused degraded comprehension performance because of the interference between disparate information perceived from visual and auditory channels. In the proposed dual-modal information presentation, the graphic and voice information are derived from the same textual information. The information should be highly relevant and complementary to each other. Therefore, the mental integration of the visual and auditory information might be easier.

With a reduced cognitive workload and easier mental integration in working memory, the proposed dual-modal information presentation may significantly improve the effectiveness of users' information comprehension.

4 Method

This study will use sample analytical tests from the Graduate Record Examination (GRE) for the experiment because these tests are designed to measure subjects' analytical comprehension and reasoning skills without assessing specific content knowledge. An experiment Web site will be built to present the GRE analytical tests. The task is to perform GRE analytical tests through the experiment Web site. Similar to the GRE analytical test, the task takes 30 minutes in our experiment.

The only independent variable is information presentation mode. There are two treatments: Text (T) mode and Graphic+Voice (GV) mode. In the T mode, all information will be visually presented as texts on a Web page. In the GV mode, the original textual information will be split into a tree-structured diagram and speech output. Three faculty members with rich teaching experience will be asked to manually convert the GRE analytical tests into a graph + voice presentation according to the proposed method (see Figure 3). Only the hierarchical relationship and related entities will be converted in graphics.

The two dependent variables are user performance and satisfaction. User performance is measured by the number of correctly answered questions within a 30-minute period. The task starts when the first analytical problem is presented on the screen, and ends when time is up. User satisfaction will be measured by a satisfaction questionnaire using a 7-point Likert scale. Based on Technology Acceptance Model (TAM) (Davis, 1989; Koufaris, 2002), this satisfaction questionnaire is designed to measure subjects' perceived usefulness and ease of use of the two interfaces. In addition, one question will be added to measure user's general satisfaction.

Sixty university students will be recruited to participate in this experiment. They will be evenly and randomly distributed into two treatment groups. Their background information will be recorded to ensure a controlled balance in demographic characteristics between groups. Because individual participants' analytical comprehension and reasoning skills may vary greatly and such skills could affect their performance in the experiment, we propose to use an independent GRE analytical test as a pre-test to estimate a participant's skills before the actual experiment task is performed. In this 15-minute pre-test, all information will be visually presented as texts on a Web page for both groups. The estimate of analytical comprehension and reasoning skills or possibly other test-taking skills from the pre-test will be used as a covariate in the analysis of the experiment task performed later.

The experiment design is a simple t-test. Subjects will perform two GRE analytical tests. The first test serves as the pre-test for estimating each individual's analytical comprehension, reasoning, and test-taking skills. The second test will be presented in T vs. GV mode for comparing the differences of these two presentation modes.

Each subject will be asked to sign a consent form before participation. During the training session, each subject will fill out a background questionnaire and the experimenter will describe the tasks included in different groups. A sample problem will be used to explain the interface, browsing rules, time limit, graphic notations (for GV-mode group), and voice control (for GV-mode group). Subjects are allowed to ask questions during the training period. They can spend as much time as they need in the training session. Subjects will be encouraged to answer as many questions as they can during the two analytical tests. They will be allowed to browse back and forth within each problem to find or correct their answers. Subjects can click a submit button to move on to the next analytical problem after they finish the current one, but they cannot go back to the previous problem. For the GV-mode presentation, pre-recorded voice information will be automatically played when the Web page is loaded on the screen. Subjects can use controls on the screen to replay voice messages. Subjects are allowed to take breaks before or after the timed tests. Upon completion of these two tests, the subject will be asked to fill out a satisfaction questionnaire. There is no time limit for this satisfaction survey. Table 1 presents the two tests and the experiment procedure.

Table 1. Experiment Tasks and Procedure

Groups	Pre-Test (15 min)	Task (30 min)	Satisfaction Survey
T-mode Group	Solve problems in T-mode presentation	Solve problems in T-mode presentation	Satisfaction questionnaire
GV-mode Group	Solve problems with T-mode presentation	Solve problems with GV-mode presentation	Satisfaction questionnaire

The following information will be saved into a database for further analysis:

- Subjects' background information.
- Subjects' answers to analytical problems in Pre-Test and Task, and time spent on each problem.
- Subjects' response to the satisfaction survey.
- Subjects' online activities (e.g., manipulating voice messages, changing answers, etc.).

5 Next stage

A controlled experiment will be conducted to test the research hypothesis and validate the proposed new dual-modal information presentation. We are currently developing experiment instruments and expect to complete the experiment during the next four months. Preliminary results of this study will be presented at the conference.

References

Anderson, R. C. & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.) *Handbook of reading research* (pp.255-292). New York: Longman.

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), 323-245.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Broadbent, D. (1958). *Perception and communication*. London: Pergamon Press.

Cellario, M. (2001). Human-centered intelligent vehicles: Toward multimodal interface integration. *IEEE intelligent systems*, July, 78-81.

Chalfonte, B., Fish, R., & Kraut, R. (1991). Expression richness: a comparison of speech and text as media for revision. In *Proceedings of CHI'91*, 21-26, Addison Wesley.

Chan, S., Fang, X., Brzezinski, J., Zhou, Y., Xu, S., & Lam, J. (2002). Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research*, 3(3), 187-199.

Cook, M. J., Cranmer, C., Finan, R., Sapeluk, A., & Milton, C. (1997). Memory load and task interference: Hidden usability issues in speech interfaces. *Engineering psychology and cognitive ergonomic*, 3, 141-150.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, Sep. 1989, 319-340.

Denis, M. (1988). Imagery and prose processing. In M. Denis, J. Engelkamp, & J. T. E. Richardson (Eds.), *Cognitive and neuropsychological approaches to mental imagery* (pp.121-132). Dordrecht / Boston / Lancaster: Martinus Nijhoff Publishers.

Dubois, M. & Vial, I. (2000). Multimedia design: the effects of relating multi-modal information. *Journal of comupter assisted learning*, 16, 157-165.

Eimer, M. (1999). Can attention be directed to opposite locations in different modalities? An ERP study. *Clinical neurophysiology*, 110, 1252-1259.

Faletti, M. V. & Wellens, A. R. (1979). From people to places: Extending a multi-modal information-processing paradigm. *Environmental psychology and nonverbal behavior*, 3(4), 248-252.

Guglielmetti, L. (2003). Standardizing automotive multimedia interfaces. *IEEE multimedia*, April, 76-78.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1989). *Mental models*. In M. I. Posner (Ed.) Foundations of cognitive science (pp.469-499), Cambridge, MA: MIT Press.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kosslyn, S. M. & Shwartz, S. P. (1977). A simulation of visual imagery. *Cognitive science*, 1, 265-295.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to onine consumer behavior. *Information systems research*, 13(2), 205-233.

Lamping, J., Rao, R., & Piroli, P. (1995). A focus + context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings the SIGCHI conference on Human factors in computing systems* (pp.401-408). Denver, Colorado, United States Pages: 401 - 408

Mayer, R. E. (1989). Models for understanding. *Review of educational research*, 59(1), 43-64.

Mayer, R. E. & Gallini, J. K. (1990). When is an illustration worth thousand words? *Journal of educational psychology*, 82(4), 715-726.

Mayer, R. E. & Anderson, R. B. (1991). Animations need narrations: an experimental test of the dual-coding hypothesis. *Journal of educational psychology*, 83(4), 484-490.

Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The psychology of computer vision* (pp.211-277), New York: McGraw-Hill.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of education psychology*, 87(2), 319-334.

Nardi, B. A., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., & Sclabassi, R. (1993). Turning away from talking heads: the use of video-as-data in neurosurgery. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp.327-334).

Navon, D. & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review*, 86(3), 214-255.

Paivio, A. (1971). *Imagery and cognitive processes*. New York: Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental representations: A dual-coding approach*. New York: Oxford University Press.

Proctor, R. & Van Zandt, T. (1994). *Human factors in simple and complex systems*. Needham Heights, MA: Allyn and Bacon.

Robertson, G. G., Mackinlay, J. D., & Card, S. K. (1991). Cone trees: Animated 3D visualizations of hierarchical information. In *Proceedings of the CHI'91 conference on human factors in computing systems* (pp.189-194).

Siewiorek, D., Smailagic, A., & Hornyak, M. (2002). Multimodal contextual car-driver interface. In *Proceedings of the fourth IEEE international conference on multimodal interfaces* (pp.367-343).

Starner, T. E. (2002). Attention, memory, and wearable interfaces. *IEEE pervasive computing*, 1(4), 88-91.

Stock, O., Strapparava, C., & Zancanaro, M. (1997). Multi-modal information exploration. *Journal of educational computing research*, 17(3), 277-185.

Streeter, L. (1998). Applying speech synthesis to user interfaces. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp.312-343), New York, NY: Elsevier Science Pub.

Titsworth, T. (2002). Telematics might steer your car into the future. *IEEE multimedia*, July, 9-10.

Treviranus, J. & Coombs, N. (2000). Bridging the digital divide in higher education. In *Proceedings of the EDUCAUSE 2000 Conference*, Nashville Tennessee.

Wickens, C. (1980). The structure of attentional resource. In R. S. Nickerson (ed.), *Attention and Performance VIII*, 239-257, Hillsdale, NJ: Lawrence Erlbaum.

Wickens, C. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (eds), *Varieties of Attention*, 63-102, New York, NY: Academic Press.

Wickens, C. D. & Ververs, P. M. (1998). Allocation of attention with head-up displays. *Technical report of aviation research lab, Institute of aviation*, DOT/FAA/AM-98/28.

Woodhead, M. M. & Baddeley, A. D. (1981). Individual differences and memory for faces, pictures, and words. *Memory & cognitive*, 9(4), 368-370.