INTL. JOURNAL OF HUMAN–COMPUTER INTERACTION, *24*(8), 1–18, 2008

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447310802537566

Taylor & Francis
Taylor & Francis Group

Development of a Dual-Modal Presentation of Texts for Small Screens

Shuang Xu, Xiaowen Fang, Jacek Brzezinski, and Susy Chan
School of Computer Science, Telecommunications, and Information Systems DePaul University, Chicago, IL

Baddeley's (1986) working memory model suggests that imagery spatial information and verbal information can be concurrently held in different subsystems. This research proposed a method to present textual information with network relationships in a "graphics + voice" format, especially for small screens. It was hypothesized that this dual-modal presentation would result in superior comprehension performance and higher acceptance than pure textual display. An experiment was carried out to test this hypothesis with analytical problems from the Graduate Record Examination. Thirty individuals participated in this experiment. The results indicate that users' performance and acceptance were improved significantly by using the "graphic + voice" presentation. The article concludes with a discussion of the implications and limitations of the findings for future research in multimodal interface design.

1. INTRODUCTION

Nowadays, most information is presented in a textual format. However, textual display may not always be the best solution. For example, people may have difficulties in reading and comprehending complex texts, users of handheld devices cannot read much text information on a small screen, and drivers cannot read texts well from a computer screen in the car. Under such circumstances, textual presentation becomes extremely inefficient. Multimodal interfaces provide a logical solution by delivering information through multiple sensory modalities.

The benefit of delivering information across different sensory modalities is often justified by the independent nature of multimodal information processing, which assumes that there will be no interference between tasks and thus no degradation in performance. However, research in cognitive psychology shows that visual and auditory perceptual processing is closely linked (Eimer, 1999). Problems related to memory and cognitive workload were found in current applications with voice-based interface (Cook, Crammer, Finan, Sapeluk, & Milton, 1997). For instance,

Correspondence should be addressed to Xiaowen Fang, 243 South Wabash Avenue, Chicago, IL 60604-2301. E-mail: xfang@cti.depaul.edu

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. Switching attention between modalities also may be slow and incur a high cost (Cook et al., 1997).

The objectives of this research are (a) to develop a dual-modal interface that improves the effectiveness of mental integration of information from different modalities, and (b) to test its effectiveness by comparing the new interface with the commonly used textual display. This study focuses on the dual-modal presentation of textual information that describes network relationships for small screens. Results of this study will help to address the usability problems associated with small-screen computers and the mobile information access via handheld devices.

2. LITERATURE REVIEW

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

2.1. Human Attention

The interference encountered during multimodal 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 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 (Kahneman, 1973) 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 (Navon & Gopher, 1979; Wickens, 1980, 1984) propose that there are several distinct subsystems, each having its own limited pool of resources. The multiple-resource models assume that two tasks can be efficiently performed together to the extent that 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 the 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 multimodal information-processing tasks, resources allocated to nondominant channel decrease, as compared to single-modal information processing. Another explanation is that mental integration of different multimodal information causes a heavy cognitive load in working memory. Performance will degrade if this integration is critical to understanding information received from different sensory channels.

Although the details of attention allocation mechanism are still under exploration, past research has positively confirmed that human's cognitive resources for attention

are relatively limited. Therefore, concurrently performing two or more tasks generally results in a drop of performance for one or all tasks. The amount of shared resources is one of the factors that determine how well people can divide their attention between tasks (Wickens, Gordon, & Liu, 1998). A large body of research work has shown that people are generally better at dividing attention across modalities, typically on visual and auditory information, than within a single modality channel (Coull & Tremblay, 2001; Dubois & Vial, 2000; Stock, Strapparava, & Zancanaro, 1997). Other researchers (Polson & Friedman, 1988; Wickens & Liu, 1988) show that imagery/spatial and verbal processing demands distinct resources, whether occurring in perception, central processing, or responding. The following literature review on working memory explains why processing spatial and verbal information concurrently does not cause competition of cognitive resources.

2.2. Working Memory

Baddeley (1986) proposes a working memory model that depicts three components: central executive, visuo-spatial sketchpad, and phonological loop (see Figure 1). 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 hold verbal and imagery information simultaneously 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 model has been confirmed by many studies. For example, Mousavi, Low, and Sweller (1995) showed that students' performance improved significantly when the verbal representation and image representation of a geometry problem were respectively presented in auditory and visual modes. They further suggested that distributing relevant information in visual and auditory modalities might effectively increase working memory.

2.3. Multimodal Interfaces

Research on interaction between sound, written words, and the image of objects shows that when different sources of information are integrated, multimedia

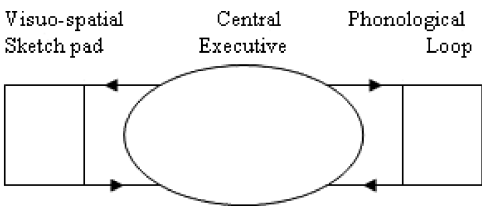


FIGURE 1 Baddeley's (1986) Working Memory Model.100

presentation can improve learners' engagement (Webster & Ho, 1997) without increasing their cognitive load (Dubois & Vial, 2000). Treviranus and Coombs (2000) demonstrated how to make the learning environment more flexible and interactive by integrating captions, video descriptions, and other access tools. Elting, Zwickel, and Malaka (2002) reported an evaluation of different modality combinations on three devices (personal digital assistant, TV set, and desktop computer). Their results indicate that the combination of spoken text in connection with a picture is the most effective regarding recall performance. This effect is strongest for users' performance on personal digital assistants, as compared to other devices.

Unlike desktop environment with large display screens, mobile devices have limited screen space to present information. Although computing performance on mobile devices has been advanced in the last decade, how to effectively present information remains a challenge for interface designers. A variety of information visualization methods have been studied and developed for desktops. However, these techniques do not apply well for mobile devices because of the inherent hardware constraints. Yoo and Cheon (2006) proposed several methods such as using a fisheye view for sequential layout and a radial view for hierarchical layout as a potential solution to improve information navigation on mobile devices. Brewster's (2002) study indicates that presenting information about buttons in sound increases their usability and allows the button size to be reduced. Therefore, more information can be presented with sonically enhanced buttons and reduced workload on mobile devices. Other researchers have been exploring systematic tools to dynamically generate multimodal presentations using output modalities determined by wireless bandwidth (Solon, McKevitt, & Curran, 2004). Paterno and Giammario (2006) provided a programming environment for designers and developers to author mobile interfaces with various combinations of graphics and voice presentations.

Although many design guidelines were developed during the empirical explorations, the complex nature of how multimodal presentations affect human's information processing has not been fully addressed. Dubois and Vial (2000) suggested that several factors affect the effectiveness of the integration of multimodal information. These factors, including the presentation mode and the construction of coreferences, relate to the different components of the learning materials and the characteristics of the task as well. Hede (2002) proposed a theoretical model of multimedia effects on learning incorporating 12 elements. Each element in this modal represents operational construct: visual input, auditory input, learner control, motivation, learner style, cognitive engagement, attention, working memory, intelligence, long-term storage, reflection, and learning. Based on the theoretical foundation of cognitive psychology, mainly in human attention and working memory, our study focuses on how to develop an effective dual-modal presentation for the given textual information.

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 graphical and verbal representations. Mayer's (1989; Mayer & Gallini, 1990; Mayer & Anderson, 1991) empirical studies show that an effective illustration model should use images or diagrams to reorganize and integrate the acquired information. The illustration must be able to guide users' selective attention toward 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 relationships among them. Dual coding theory (Paivio, 1971, 1986) further predicts that concrete language could be better comprehended and more easily integrated into memory than abstract language because two forms of mental representation, verbal and imagery, are available for processing concrete information.

Schemas (or scripts, frames) have been widely used in knowledge representation (Johnson-Laird, 1983, 1989; Proctor & Van Zandt, 1994). 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 (Proctor & Van Zandt, 1994). Knowledge is often represented as a network of interrelated units. Such networks are commonly presented to the user as a diagram of nodes connected by lines. These diagrams have provided a powerful visual metaphor for knowledge representation. Travers's (1989) study indicates that when this type of diagrammatic representation of knowledge structures matches people's virtual metaphor, it improves their information comprehension.

Based on these discussions, the authors propose a dual-modal information presentation that presents the internal relationship depicted in texts as network 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

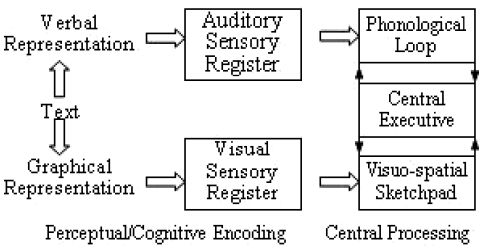


FIGURE 2 Splitting textual information.

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 central executive for comprehension.

In this proposed dual-modal presentation (see Figure 3), network relationships contained in texts will be extracted and presented in diagrams. The remaining textual information will be delivered through the auditory channel. As the Technology Acceptance Model (TAM; Davis, 1989; Venkatesh & Davis, 2000) indicates, perceived usefulness and perceived ease of use are the two determinants of a user's intention to adopt a new technology. The following hypotheses are proposed to test the effectiveness and user acceptance of this dual-modal information presentation.

H1: The dual-modal presentation of network relationships will result in superior comprehension performance as compared to pure textual display.
H2: The perceived ease of use of the dual-modal presentation of network relationships will be greater than that of pure textual display.
H3: The perceived usefulness of the dual-modal presentation of network relationships will be greater than that of pure textual display.

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. A reduced cognitive workload is expected by concurrently utilizing the two subsystems in working memory to process the same amount of information. Research in human attention shows that many voice-based interfaces caused degraded comprehension performance because of the interference of 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 and should be highly relevant and complementary to each other. Therefore, the mental integration of the visual and auditory information will be easier during comprehension.

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. Because of the reduced difficulty

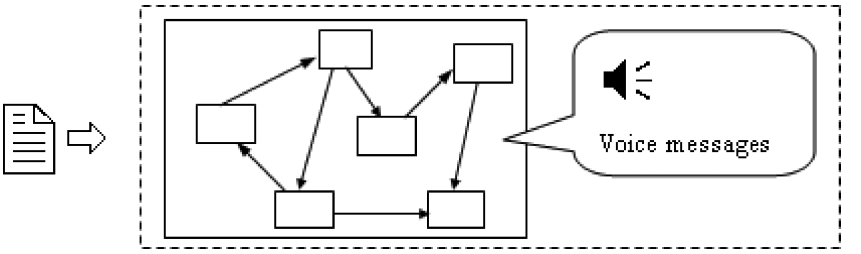


FIGURE 3 Proposed dual-modal presentation.

of mental integration and improved effectiveness, users may find the dual-modal presentation with greater perceived ease of use and perceived usefulness.

4. METHOD

This study used sample analytical tests from the Graduate Record Examination (GRE) for the experiment because these tests were designed to measure participants' analytical comprehension and reasoning skills without assessing specific content knowledge. An experiment Web site was built to present the GRE analytical tests. The task was to perform GRE analytical tests through this Web site. The following sections discuss the participants, the tasks and experiment system, independent and dependent variables, and the procedure.

4.1. Participants

Thirty participants were recruited from a Midwest university in the United States to participate in this study. The sample was composed of undergraduate students, graduate students, staff, faculty members, and alumni. Representing a wide range of background in terms of age, ethnicity, computer experience level, and Internet usage, these participants were recruited via ads placed on the university portal site. Participants were randomly assigned to one of the two groups: textual display (T-mode) group and "graphics + voice" display (GV-mode) group, with a controlled balance in gender and native language. Participants' background information was collected with a questionnaire before pretest (see Table 1). Considering the possible interplay among demographic factors and users' performance or acceptance of the new interface, a correlation analysis was carried out and confirmed that participants' demographic characteristics were balanced between the two groups and did not bias the findings of this study.

4.2. Experiment Design and Tasks

The experiment design was a between-subject simple *T* test. Participants performed two GRE analytical tests. The first test served as the pretest for estimating each participant's analytical comprehension, reasoning, and test-taking skills. Both groups went through the same pretest in textual presentation. The second test, serving as the task in this experiment, was presented in T-mode versus GV-mode, respectively, for the two groups in order to compare the differences of these two presentation modes. An illustration of the two presentation modes is shown in Figure 4.

The task was to answer multiple-choice questions based on the description of analytical problems from GRE sample tests. Participants were encouraged to answer as many questions as they could in the two analytical tests. Each analytical problem was presented on one framed HTML page, with problem description displayed in the upper frame (nonscrollable) and questions listed in the lower frame (scrollable). This interface was designed to ensure that the presentation of each problem was visible during the completion of each problem. Participants could

Table 1: Demographic Information of Participants

	Group	
Demographic Characteristics	*T-Mode*	*GV-Mode*
Gender		
Male	6	6
Female	9	9
Language		
Native English speakers	13	13
Non-Native English speakers	2	2
Average age		
19–26 years old	7	7
27–36 years old	4	5
37–46 years old	3	2
47+ years old	1	1
Educational background		
High school	0	0
Undergraduate	7	7
Graduate	8	7
Ph.D.	0	1
Visually or aurally impaired	0	0
Internet Usage		
0–4 years	0	0
5–10 years	13	12
11+ years	2	3
Frequency of Internet usage		
Hourly	3	6
Daily	12	9
Computer applications with VA output		
Never used before	4	4
Used before	11	11
Usage of VA applications		
1–4 years	4	2
5–9 years	5	8
10+ years	2	1
Frequency of VA application use		
Hourly	2	1
Daily	1	0
Weekly	4	3
Monthly	4	7
GRE		
Never took before	12	13
Took before	3	2
Last time took GRE		
1 year ago	2	0
10 years ago	0	1
20 years ago	1	1

Note. VA = visual-auditory; GRE = Graduate Record Examination.

click on the Submit button to move on to the next problem after they finished the current one, but they could not go back to the previous problem. For the T-mode presentation, analytical problems were presented in plain text. For the GV-mode presentation, prerecorded voice information was played automatically as the Web

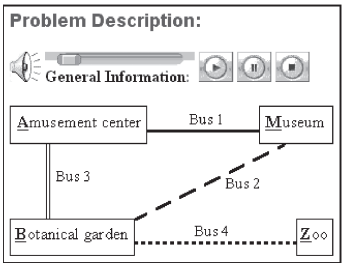
T-Mode Information Presentation

Problem Description:

Central Park has a railcar system that transports visitors to different stops in the park. There are four buses, numbered 1 through 4, which serve the 4 stops - the museum, the botanical garden, the amusement center, and the zoo - in the following way:

- Bus 1 travels between the amusement center and the museum.
- Bus 2 travels between the botanical garden and the museum.
- Bus 3 travels between the amusement center and the botanical garden.
- Bus 4 travels between the botanical garden and the zoo.

GV-Mode Information Presentation



Voice message:

"Central Park has a railcar system that transports visitors to different stops in the park. There are four

Sample questions:

1. A trip using each of the buses exactly once will frequent each of the spots exactly once if it begins at which spot and ends at which spot?
   (A) It begins at M and ends at Z. (B) It begins at Z and ends at M.
   (C) It begins at B and ends at Z. (D) It begins at A and ends at M.
   (E) It cannot be done as stated.

2. If a family takes each bus exactly once, which of the following is a complete and accurate list of the stops where they must have stopped exactly twice?
   (A) botanical garden (B) amusement center and botanical garden
   (C) botanical garden and museum (D) botanical garden and zoo
   (E) botanical garden, museum, and zoo

3. Which one of the following sequences of buses is NOT possible?
   (A) 1 to 3 to 2 to 4 to 3 (B) 2 to 1 to 3 to 2 to 1
   (C) 3 to 4 to 4 to 2 to 1 (D) 4 to 2 to 2 to 4 to 4
   (E) 4 to 3 to 1 to 2 to 4

4. To go from the zoo to the amusement center in the fewest number of stops requires how many buses?
   (A) 1 (B) 2 (C)3 (D) 4 (E) 5

FIGURE 4 An illustration of presentation modes.

page was loaded on the screen. Participants could use buttons to "play," "pause," or "stop" the voice messages. A voice message would be stopped automatically if another voice message was activated to ensure that only one voice message was playing at any given time. Participants were allowed to take breaks before or after the two tests. Participants were not allowed to use paper and pen so that both groups could only perceive information as it was presented in the experiment.

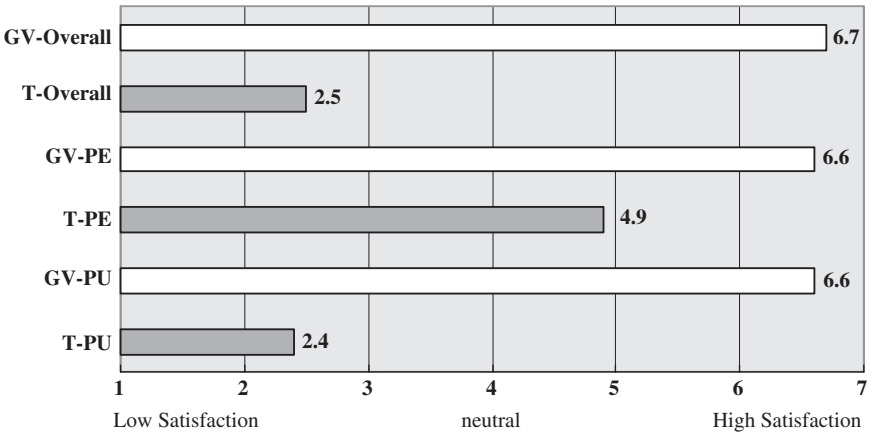


FIGURE 5 Comparison of average user acceptance between groups.

4.3. Independent and Dependent Variables

The only independent variable was information presentation mode. There were two treatments: T-mode and GV-mode. In the T-mode, all information was visually presented as texts on a Web page. In the GV-mode, the original textual information was split into speech output and network graphs that represented the relationship according to the proposed method (see Figure 4). Only the internal relationship and related entities were converted into graphics. Figure 4 provides the screenshots of an example of the T-mode and GV-mode of information presentation.

The three dependent variables were users' performance, perceived ease of use, and perceived usefulness. Users' performance was measured by the number of correctly answered questions within a task session. A task session started when the first analytical problem was presented on the screen and ended after 30 min. For each problem in the task session, participants viewed the description of this problem and answered questions related to this problem, then moved to the next problem. A sufficient number of problems were included to make sure that no participant could finish all of the problems before the end of the task session. Perceived ease of use and perceived usefulness were measured by a questionnaire using a 7-point Likert scale, from 1 (*strongly disagree*) to 7 (*strongly agree*). As the TAM (Davis, 1989; Koufaris, 2002; Venkatesh, 2000; Venkatesh & Davis, 2000) indicates, perceived usefulness and perceived ease of use are of primary relevance for information technology adoption behaviors. In TAM, perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance," and perceived ease of use is defined as "the degree to which a person believes that using a particular system would be free of effort." Among the 13 statements in this questionnaire, 6 measure the perceived usefulness during the task (e.g., "The presentation of the problems helped me find answers quickly"), and another 7 statements measure the perceived ease of use (e.g., "The presentation of the problems was clear and understandable"). See Table 2.

Table 2: Independent Variable and Dependent Variables

Independent variable	Information presentation mode: Text mode vs. Graph + Voice mode
Dependent variables	1. Performance: Number of correctly answered questions within 30 min
	2. Perceived ease of use
	3. Perceived usefulness

4.4. Procedure

This experiment was conducted for participants individually on one computer, one at a time. Each participant submitted an online consent form before participating in the experiment. After completing a background questionnaire, the participant was assigned to either the T-mode or GV-mode group with controlled balance of gender and language. The experimenter then described the task using a sample analytical problem to explain the interface, browsing rules, and time limit. In addition, instructions regarding how to understand graphic notations and how to use voice control were provided for participants in the GV-mode group. Participants were allowed to ask questions during this training session. They could spend as much time as they needed during the training. Next, each participant started the pretest and then the task session. Upon completion of the task session, each participant was asked to fill out an acceptance questionnaire. There was no time limit for this acceptance survey. Finally, each participant was debriefed about his or her task experience and the interface used in the experiment.

5. RESULTS AND DISCUSSION

5.1. Results

Performance. The hypotheses state that the dual-modal presentation will result in superior comprehension performance and greater user acceptance, as compared to the traditional textual presentation. The result of task performance, measured by the number of correctly answered questions in the experiment tasks, is listed in Table 3.

To alleviate the impact of individual differences, an analysis of covariance was conducted using participants' pretest performance as a covariate to adjust the results of their task performance. Results of the analysis of covariance reveal a

Table 3: Descriptive Statistics of Task Performances

	Pretest				*Task*			
	T-Mode		*GV-Mode*		*T-Mode*		*GV-Mode*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
No. of correct answers	4.4	1.35	4.9	2.84	9.5	3.02	23.1	8.55
Accuracy	0.617	0.214	0.466	0.229	0.487	0.1663	0.757	0.1794

significant difference ($F = 92.50$, $p < .0001$) in the task performance between the T-mode group and GV-mode group. As shown in Table 3 (T-mode: $M = 9.5$, GV-mode: $M = 23.1$), the number of correct answers on average increased by more than 143% ($23.1/9.5 \approx 2.432$) for the GV-group who used the dual-modal interface.

Accuracy, defined as the number of correctly answered questions divided by the total number of answered questions, was calculated for precaution because if a participant guessed many times during the experiment, he or she might be able to correctly answer more questions with a much lower accuracy. A significant difference ($F = 27.05$, $p < .0001$) was found in term of task accuracy between the two groups (T-mode: $M = 0.487$, GV-mode: $M = 0.757$). The proposed dual-modal information presentation resulted in a significant improvement in participants' task accuracy by more than 55% ($0.757/0.487 = 1.5544$). Because the probability of hitting the correct answer would be 20% (five choices for each question), the significant difference in accuracy (> 55%) indicates that the improvement in the GV-mode group was not a result of random guessing. These results confirm that participants in the GV-mode group did perceive and process information more effectively.

Perceived ease of use and perceived usefulness. Perceived ease of use and perceived usefulness were measured as surrogates of satisfaction. A factor analysis with varimax rotation was performed to establish convergent and discriminant validity of the two constructs. As a result of low loadings in the factor analysis, one item was removed from the original six statements measuring perceived usefulness, and four items were removed from the original seven statements measuring perceived ease of use. Cronbach's alpha values were calculated to verify reliability of the instrument. The high Cronbach's alpha values for perceived usefulness ($\alpha = 0.98$) and perceived ease of use ($\alpha = 0.86$) suggested that the questionnaire was reliable and valid.

After removing the five items from the questionnaire, the total scores of the items measuring perceived ease of use and perceived usefulness were respectively calculated and used in the analysis. *T* tests were conducted to compare the differences between the two groups in perceived usefulness and perceived ease of use. Transformation was conducted to satisfy the homogeneity of variance assumption of the *T* test.

As shown in Table 4, results of *T* tests indicate significant differences between the GV-mode and T-mode in perceived usefulness, $t(28) = 13.38$, $p < .0001$, and perceived ease of use, $t(28) = 3.99$, $p = .0004$. Users' scores of perceived usefulness (T-mode: $M = 12.1$ and GV-mode: $M = 33.1$) were the sum of five items used in the questionnaire. Therefore, the average score of perceived usefulness in the T-mode

Table 4: Descriptive Statistics and *T*-Test Results of Acceptance

	T-Mode[a]		*GV-Mode*[b]			
	M	*SD*	*M*	*SD*	*t*	*Pr* > | *t* |
Perceived usefulness	12.1	5.46	33.1	2.72	13.38	< .0001
Perceived ease of use	14.6	4.85	19.9	1.79	3.99	.0004

[a]$n = 15$. [b]$n = 15$.

group was 12.1/5 = 2.4, and 33.1/5 = 6.6 in the GV-mode group. Three items were used to measure users' perceived ease of use, the average score of perceived ease of use in the T-mode group was 14.6/3 = 4.9, and 19.9/3 = 6.6 in the GV-mode group. These findings confirm that participants in the GV-mode group found the proposed dual-modal interface more acceptable.

5.2. Discussion

Results of this experiment fully supported the three hypotheses H1, H2, and H3. A significant improvement in task performance and users' acceptance were found in the GV-mode group. Major findings in this experiment are summarized here:

1. GV-mode participants' average number of correct answers increased by more than 143%, as compared to the task performance of their counterparts in the T-mode group.
2. Participants' average accuracy on the GV-mode presentation increased by more than 55%, in contrast to the task accuracy on the T-mode presentation.
3. Participants in the GV-mode group expressed much higher acceptance of the interface as measured by perceived usefulness and perceived ease of use. This finding provides subjective evidence that users could perceive complementary information from visual and speech representations.
4. Many participants in the GV-mode group felt that the graphic illustration was very helpful and the diagrams seemed to have contributed more during the information comprehension in the experiment.

These results bolster prior theories in human attention and working memory. As suggested by Wickens's (1980, 1984) multiple resource pool model, two or more tasks can be performed together efficiently to the extent that they require separate attentional resources, such as in visual and auditory modalities. Baddeley's (1986) working memory model also predicts that tasks concurrently utilizing different subsystems (visuo-spatial sketch pad and phonologic loop) in the working memory should have minimum interference. Although users are browsing imagery or graphical information, they may be able to receive verbal information from the auditory channel. The results agree with Paivio's (1986) dual coding theory as well, which suggests that the association at a perceptual level with image coding and the association at a semantic level with verbal coding will enhance the interconnections between the dual-modal representations. The materials used in the proposed dual-modal presentation required users' cognitive coordination between the auditory and visual components because information presented in each modality alone was insufficient for problem solving. This coordination, or mental integration, is assumed to be governed by the central executive (Baddeley, 1992), which has been supported by many other studies. Some researchers (Logie, Gilhooly, & Wynn, 1994) have found that disruption of the central executive could severely degrade working memory performance. Yee, Hunt, and Pellegrino (1991) revealed that the ability to coordinate perceptual and verbal information was different from processing perceptual and verbal performance alone.

In this experiment, information from the diagrams and voice messages must be integrated before the participants would completely understand the description of each problem. The relatively loose association between the two presentation components could demand more working memory resources for the coordinating function performed by the central executive. However, as suggested by other researchers (Mousavi et al., 1995), it is possible that more working memory resources are available for coordination when distinct but related information is distributed simultaneously in auditory and visual subprocessor rather than in either one alone.

The findings also concur with the results of Mayer's (1989; Mayer & Gallini, 1990) in which an illustration becomes more effective when images or diagrams are reorganized and integrated to present the relationship among the acquired information entities. Feedback from the T-mode participants indicated that users did not have problem understanding the given information. However, it was difficult to mentally integrate the information cues without using additional aids, such as paper and pencil, to temporarily store information. The graphic integration of network relationships might have reduced the cognitive workload required for schema acquisition and automation in working memory (Johnson-Laird, 1983, 1989; Proctor & Van Zandt, 1994). As compared to the auditory presentation, the graphic presentation might have contributed more to participants' comprehension performance in the GV-mode.

In summary, presenting the textual description of network relationships as graphics with relevant but additional speech output has significantly improved users' comprehension performance as well as their acceptance of the interface. All three hypotheses were fully supported by the results of this experiment.

6. CONCLUSIONS

In this study, the authors aimed to improve the effectiveness of presenting information using multiple modalities. A dual-modal information presentation was proposed, developed, and tested through a controlled experiment. Findings from this study suggest the following:

1. Users could concurrently integrate perceived visual (diagrams) and auditory (voice messages) input with an improved efficiency.
2. Highly relevant speech information might facilitate users' understanding of diagrammatic information.
3. The distribution of cognitive workload across modalities may reduce demand for mental efforts required for information comprehension, as compared to single-modality presentation.
4. Users might report higher acceptance of a dual-modal interface because of the reduced mental workload during information processing.

6.1. Contributions

The findings of this study have potential implications for future research in multimodal interface design. Multimodal interfaces are specially promising to mobile

applications because of the nature of wireless technology. Mobile devices have two main constraints: small screen size and their mobile usage (Chan et al., 2002). Compared to desktop or laptop computers, mobile devices typically have a very small screen for information display. When the device is used on the move, it makes reading textual information much more difficult. Multimodal interfaces could help to address these constraints by delivering information through multiple sensory modalities such as visual and auditory channels. The advancement of speech synthesis technology provides a strong support for information delivery via the auditory channel. Meanwhile, the amount of texts could greatly be reduced after being converted into diagrams. An increased readability with a decreased requirement of screen real estate is expected in the visual/auditory presentation.

The contribution of this research is beyond the interface design of handheld devices. Results of this study can facilitate generic design of human–computer interaction as well as instructional information presentation. Because visual and auditory perceptual processing is closely linked, perception of disparate information from different modality channels often introduces interference and distraction. The mental integration of the dual-model information also causes a heavy cognitive memory load. Therefore, degraded performances have been found in current computer-based applications with visual-auditory output. Researchers have spent years exploring different aspects of how to efficiently utilize human sensory modalities. Results of this experiment indicate that when information is converted into a visual graphic + auditory speech presentation, it is likely to simultaneously use the visuo-spatial sketchpad and the phonological loop in the working memory system. In the proposed dual-modal presentation, the graphic and speech information are derived from the same textual content. Thus the mental integration of the highly relevant and compatible visual and auditory information does not demand a high cognitive workload.

6.2. Limitations and Future Research Directions

One major limitation of this study is that it did not separate the effects of graphics from those of voice. A future study comparing textual display, "Graphics + Voice," and "Graphics + Text" would be necessary to further investigate how graphics and voice affect user performance and satisfaction respectively. The primary goal of this study was to use working memory theories to develop an effective dual-modal interface for small screens, with distinct but compatible information presented in different modalities. The results have fulfilled this goal. Although this experiment was not designed to compare the contributions of graphic and auditory presentations to users' information comprehension, the possibility of alternative interpretations of these results should not be discounted. For instances, was users' comprehension performance improved mainly because the clues were integrated in the diagrams? Which component, the voice messages or the graphics, played a more active role in affecting users' comprehension in the proposed dual-modal interface? A future study with a factorial design will provide a better understanding of the aforementioned research questions.

Other limitations should also be considered when interpreting the results. This experiment explored specific textual information that explicitly describes network relationships. This might limit the scope of the application of the findings previously discussed. As suggested by many researchers in the multimodal information presentation (Dubois & Vial, 2000; Hede, 2002), the effectiveness of a multimodal information presentation can be affected by several factors such as the presentation mode, the content of the presented information, as well as the characteristics of the performed task. Sample problems taken from the GRE tests were used in the experiments to measure participants' analytical comprehension and reasoning skills without assessing specific content knowledge. However, caution should be exercised when applying findings from this study because the results may not hold true when the content or the nature of the information-processing task changes.

In the future, the authors will continue to explore different methods of converting texts to graphics. Advanced natural language processing techniques will be employed to generate heuristics for text-graphics conversion and speech synthesis. Such process should eventually be automated and applicable to a wider range of textual information. Future research will likely enjoy a greater degree of success if subjective measures of cognitive load and the efficiency assessment of individual presentation components can be incorporated. Although the participants of this study represented a variety of background in terms of age, ethnicity, computer experience level, and Internet usage, they were all recruited from one university. People who have different occupations and a variety of educational backgrounds should also be observed in future studies to explore possible different behavior patterns when they perform comprehension tasks on the proposed dual-modal interface.

REFERENCES

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

Baddeley, A. D. (1992). Working memory. *Science*, *225*, 556–559.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing*, *6*(3), 188–205.

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

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.

Coull, J., & Tremblay, L. E. (2001). Examining the specificity of practice hypothesis: Is learning modality specific? *Research Quarterly for Exercise & Sport*, *72*(4), 345–354.

Curtis, R. V. (1988). When is a science analogy like a social studies analogy? A comparison of text analogies across two disciplines. *Instructional Science*, *13*, 169–177.

Davis, F. D. (1989, September). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, pp. 319–340.

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.

Elting, C., Zwickel, J., & Malaka, R. (2002). Device-dependant modality selection for user-interfaces: An empirical study. In *Proceedings of the 7th International Conference on Intelligent User Interfaces IUI '02* (pp. 55–62). New York: ACM Press.

Hede, A. (2002). An integrated model of multimedia effects on learning. *Journal of Educational Multimedia and Hypermedia*, *11*(2), 177–191.

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.

Logie, R., Gilhooly, K., & Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition*, *22*, 395–410.

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research*, *59*(1), 43–64.

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.

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

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.

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.

Paternò, F., & Giammarino, F. (2006). Authoring interfaces with combined use of graphics and voice for both stationary and mobile devices. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 329–335). New York: ACM Press.

Polson, M. C., & Friedman, A. (1988). Task-sharing within and between hemispheres: a multiple-resources approach. *Human Factors*, *30*(5), 633–643.

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

Solon, A., McKevitt, P., & Curran, K. (2004). Mobile multimodal presentation. In *Proceedings of the 12th annual ACM international conference on Multimedia table of contents* (pp. 440–443). New York: ACM Press.

Stock, O., Strapparava, C., & Zancanaro, M. (1997). Multi-modal information exploration. *Journal of Educational Computing Research*, *17*(3), 277–185.

Travers, M. (1989). A visual representation for knowledge structures. In *Proceedings of the Second Annual ACM Conference On Hypertext* (pp. 147–158). New York: ACM Press.

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

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Webster, J., & Ho, H. (1997). Audience engagement in multimedia presentations. *ACM SIGMIS Database*, *28*(2).

Wickens, C. (1980). The structure of attentional resource. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 239–257). Hillsdale, NJ: Erlbaum.

Wickens, C. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–102). New York: Academic Press.

Wickens, C. D., Gordon, S. E., & Liu, Y. (1998). *An introduction to human factors engineering*. New York: Addison Wesley Longman.

Wickens, C. D., & Liu, Y. (1988). Code and modalities in multiple resources: A success and a qualification. *Human Factors*, *30*(5), 599–616.

Yee, P., Hunt, E., & Pellegrino, J. (1991). Coordinating cognitive information: Task effects and individual differences in integrating from several sources. *Cognitive Psychology*, *23*, 615–680.

Yoo, H. Y., & Cheon, S. H. (2006). Visualization by information type on mobile device. In *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualization* (Vol. 60, pp. 143–146). Darlinghurst, Australia: Australian Computer Society, Inc.

APPENDIX

Measurement Validation

Table 1: Factor Analysis

	Factor 1 (Perceived Usefulness)	*Factor 2 (Perceived Ease of Use)*	*Item Communality*
Question 1	0.85		0.85
Question 3	0.85		0.91
Question 5	0.91		0.97
Question 9	0.79		0.84
Question 13	0.91		0.94
Question 2		0.81	0.77
Question 4		0.62	0.45
Question 10		0.80	0.82

Note. Items Q6, Q7, Q8, Q11, and Q12 were dropped because of their low loadings.

Table 2: Internal Consistency of the Instrument

Variable	*No. of Items*	*Cronbach's α Value*
Perceived usefulness	5: Q1,Q3,Q5,Q9,Q13	0.98
Perceived ease of use	3: Q2, Q4, Q10	0.86