Development of a dual-modal information presentation of sequential relationship

1. Introduction

Technologies for speech recognition and synthesis are becoming increasingly sophisticated and provide support for information processing via multi-modal interfaces. The benefit of delivering information across different sensory modalities is often backed by the independent nature of multi-modal information processing, which assumes that there will be no interference of modality between tasks and thus no degradation in performance (Cook *et al*. 1997). Research in cognitive psychology shows that visual and auditory perceptual processing is closely linked (Eimer 1999). However, problems related to memory and cognitive workload are found in applications with voice-based interface (Cook *et al*. 1997). For instance, mental integration of disparate information from different modality channels causes a heavy cognitive workload. As transient auditory information, speech presentation may impose a greater memory burden. Also, switching attention between modalities may be slow and have a high cognitive cost.

The objective of this research is to design dual-modal interfaces 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. This study focuses on the dual-modal presentation of textual information that describes sequential relationships or chronological events. The proposed dual-modal presentation of the selected textual information describes such relationships among entities in diagrams and delivers the remaining information as voice messages. Users are expected to have superior comprehension performance and higher satisfaction while using the graph + voice presentation than using the pure textual presentation.

2. Background literature

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

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. Previous research and theories in human attention and allocation of attentional resources are summarized in table 1.

Table 1. Research in Human Attention

Author(s)	Theme	Findings/Propositions
Broadbent (1958)	Attention model	Bottleneck models suggest that only a limited amount of information can be brought from the sensory register to working memory.
Kahneman (1973)	Attention model	Resource models view attention as a limited-capacity resource that can be allocated to one or more tasks.

Navon and Gopher (1979); Wickens, (1980 and 1984)	Attention model	Multiple-resource models propose that there are several distinct subsystems, each having their own limited pool of resources. Therefore, two tasks can be efficiently performed together to the extent that they require separate pools of resources.
Cook *et al*. (1997)	Multimodality	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. (e.g. dual-modal interface used in prototype of cockpits.)
Wickens, Gordon and Liu (1998)	Multimodality	The amount of shared resources affects how well people can divide their attention between tasks.
Dubois and Vial (2000); Treviranus and Coombs (2000); Coull and Tremblay (2001); Tremblay and Proteau (1998); Stock, Strapparava and Zancanaro (1997)	Multimodality	People are generally better at dividing attention cross-modality, typically on visual and auditory information, as compared to processing distinct information presented within a single modality channel.
Polson and Friedman (1988); Wickens and Liu (1988)	Multimodality	Imagery/spatial and verbal processing demand distinct resources, whether occurring in the perception, central processing, or responding stage of the information processing.

2.2. Working memory

Baddeley (1986) proposes a working memory model that depicts the relationships among three components: central executive, visuo-spatial sketchpad, and phonological loop (see figure 1).

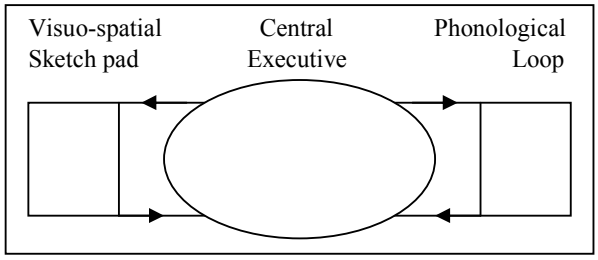


Figure 1. Baddeley's Working Memory Model (1986)

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 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) show that students' performance was significantly improved when the verbal and image representations of a geometry problem were respectively presented in auditory and visual modes. They further suggest that distributing relevant information in visual and auditory modalities might effectively increase working memory.

2.3. Visual and auditory information presentation

Introducing voice into the traditional visual interface presents new challenges. Both the nature and user perception of the signals from different sensory modalities may affect user comprehension. Previous research in visual and auditory information presentation is summarized in table 2.

Table 2. Research in Visual and Auditory Information Presentation

Author(s)	Modalities of the Interface	Findings
Archer, Head, Wollersheim and Yuan (1996)	Text and speech	After comparing the effectiveness of information delivery in visual, auditory, and visual-auditory modes, the authors 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.
Baggett and Ehrenfeucht (1983)	Visual and auditory	Three presentations were studied: visual and narration presented simultaneously, visual followed by narration, and narration followed by visual. Results suggest that there is no competition for resources when related information is presented simultaneously in visual and auditory channels.
Chalfonte, Fish and Kraut (1991)	Text and speech	Voice is more informal and interactive for handling the complex, equivocal and emotional aspects of collaborative tasks. Voice was preferred for addressing higher level issues in suggesting document modifications, but text was preferred for more detailed and lower level comments.
Cohen (1992)	Keyboard, pointing device, screen.	Users need to utilize at least two channels such as auditory and keyboard to complete a task.
DeHaemer and Wallace (1992)	Visual and auditory	Duplicate instructions were added as voice to a microcomputer workstation for decision support. The visual and audio modes of receiving information appear to be non-interfering.
Dubois and Vial (2000)	Text, image, and speech	The integration of sound, written words, and the image of the verbal information results in a light cognitive overload, which improves the effectiveness of learning.
Nardi *et al*. (1993)	Video, speech, and numerical data.	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.
Nugent (1982)	Pictures, audio, and print.	Student learning could be improved when the same content was presented in all three channels (picture, audio, and print). When different information was presented in the visual and audio modes, however, student learning was not affected by the addition of new channels and the presence of visual information did not interfere with processing the audio and vice versa.
Proctor and Van Zandt (1994)	Visual vs. auditory displays	Spatial information is best conveyed through visual displays because spatial discrimination can be made most accurately with vision. Auditory displays work best with temporal information because temporal organization is a primary attribute of auditory perception.

Schlosser, Belfiore and Nigam (1995)	Speech	The presentation of additional auditory stimuli in the form of synthetic speech is effective in assisting individuals with mental retardation to learn associations between graphic symbols with spoken words.
Sipior and Garrity (1992)	Visual and auditory	Presentation with a mix of audio and visual accompaniments improve receptiveness attributes such as perception, attention, comprehension, and retention.
Streeter (1998)	Speech and text	The main advantage of using speech is that it can be universally accessed by everyone and on the move. One notable disadvantage is that voice delivers information at a slower rate than text. Any combination of voice and text is likely to slow information acquisition process.
Treviranus and Coombs (2000)	Visual, auditory, and captioning.	The integration of captioning, video description, and other access tools for interactive information exploration makes the learning environment more flexible and engaging for students.

As suggested by Dubois and Vial (2000), 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 task characteristics. They emphasized that caution should be exercised when applying findings from multimodal information presentation studies, because these findings might not hold true when the content or the nature of the information processing task changes.

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 split into imagery/graphical and verbal representations. Table 3 summarizes prior research on graphical representation of texts.

Table 3. Research in Graphical Representation of Texts

Author(s)	Findings
Denis (1988)	Narrative texts that strongly elicit visual imagery for characters, scenery, and events are highly imageable.
Mayer (1989); Mayer and Gallini (1990); Mayer and Anderson (1991)	The illustration must be able to guide user's selective attention towards the key items in the presented information. These key items include the major entities and the relationships among them.
Paivio (1986 and 1971)	Dual coding theory predicts that concrete language should be better integrated in memory and comprehended than abstract language because two forms of mental representation, verbal and imagery, are available for processing concrete information.
Proctor and Van Zandt (1994); Johnson-Laird (1983 and 1989)	Schemas represent the structure of a person's interest and knowledge, which enables a person to develop the expectancy about what will occur.
Travers (1989)	Knowledge-based systems were often represented as a network diagram of nodes connected by lines. This representation provides a powerful visual metaphor. When this diagrammatic representation of knowledge structures matches people's virtual metaphor, it improves their information comprehension.

Denis' (1988) finding suggests that sequential information contained in texts can be converted into imagery. Such imageries may help users form schemas by reducing the cognitive demand on them for converting textual information into effective schemas. Therefore, user comprehension of the information may improve because the imagery information can be processed by the visual-spatial sketchpad (Baddeley 1986). Knowledge-based information systems are often represented as diagrams of interrelated units connected by lines. Travers (1989) indicates that when the diagrammatic representation of knowledge structures matches people's virtual metaphor, it improves their information comprehension.

Based on the above discussions, we propose a dual-modal information presentation that presents the sequential information contained in texts as flowchart-like diagrams, and delivers the remaining textual information as voice message. The following section discusses this dual-modal presentation in detail.

3. Proposed dual-modal information presentation

Based on Baddeley's (1986) working memory model, the effectiveness of human information processing could improve if the verbal and the imagery/graphical representations split from same textual information are presented via auditory and visual output, respectively. As shown in figure 2, if the verbal representation of the partial textual information is presented via auditory channel, it 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 the 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.

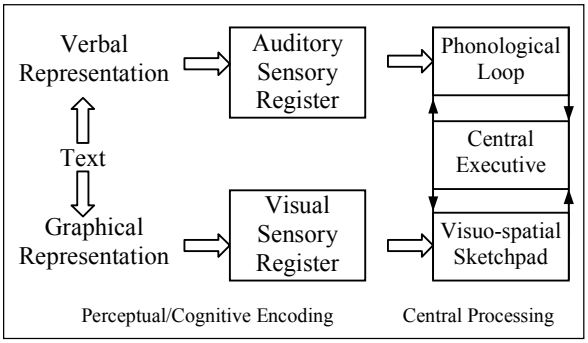


Figure 2. Splitting Textual Information

In this proposed dual-modal presentation (see figure 3), sequential relationships contained in texts are extracted and presented in diagrams. The remaining textual information is delivered through the auditory channel. The following hypothesis is proposed to test the effectiveness of this dual-modal information presentation.

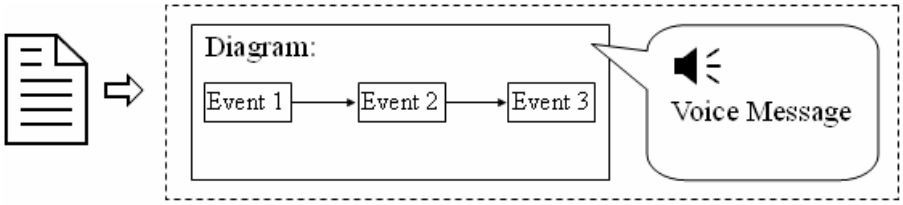


Figure 3. Proposed Dual-modal Presentation for Sequential Relationships

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

According to Baddeley's model, the 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 into speech input and temporarily saved in the phonological loop of working memory before further processing (see figure 4).

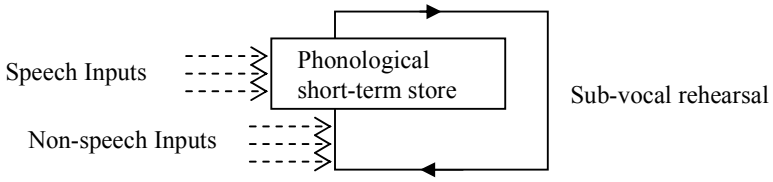


Figure 4. Structure of Phonological Loop

In the proposed dual-modal presentation, the graphical information might be perceived and held in the visuo-spatial sketchpad while the speech input is received and directly stored in the 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 shows that many voice-based interfaces have caused degraded comprehension performance because of the interference between disparate information perceived from visual and auditory channels. In the proposed dual-modal information presentation, graphic and voice information are derived from the same textual information, and should be highly relevant and complementary to each other. Therefore, 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 presentations may significantly improve the effectiveness of user information comprehension.

4. Method

This study used 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 was built to present the GRE analytical tests, both for pre-test and experiment task. The following sections discuss the subjects, the tasks and experiment system, independent and dependent variables, and the procedure.

4.1. Subjects

Thirty participants were recruited from a Midwest university in the United States. The sample was composed of undergraduate students, graduate students, staff, faculty members, and alumni. These participants represented a wide range of background in terms of age, ethnicity, level of computer experience, and Internet usage. Participants were randomly assigned to one of the two groups: textual display (T-mode) group and 'graphics + voice' display (GV-mode) group. As shown in table 4, participants in both groups shared similar profiles.

Table 4. Demographic Information of Participants

Demographic characteristics		Group	
		T Mode	GV Mode
Gender	Male	7	7
	Female	8	8
Language	Native English Speakers	10	10
	Non-Native English Speakers	5	5
Average age	19~24 years old	3	5
	25~34 years old	10	8
	35~44 years old	2	1
	45~55 years old	0	1
Educational background	High School	0	0
	Undergraduate	6	8
	Graduate	8	6
	Ph.D.	1	1
Visually or aurally impaired		0	0
Internet Usage	0-5 years	1	1
	6-10 years	12	12
	11-15 years	1	2
	More than 15 years	1	0
Frequency of Internet usage	Hourly	8	5
	Daily	7	10
Computer applications with visual-auditory (VA) output	Never used before	1	3
	Used before	14	12
Usage of VA applications	1-4 years	4	4
	5-9 years	5	5
	10-15 years	4	3
	More than 15 years	1	0
Frequency of VA application use	Hourly	2	1
	Daily	3	5
	Weekly	3	3
	Monthly	6	3
Graduate Record Examination	Never took before	11	13
	Took before	4	2
Last time took GRE	1 year ago	1	1
	2.5 years ago	1	0
	5+ years ago	2	1

4.2. Experiment design and the tasks

The experiment design was a between-subject simple t-test. Subjects performed two GRE analytical tests. The first test served as the pre-test for estimating each individual's analytical

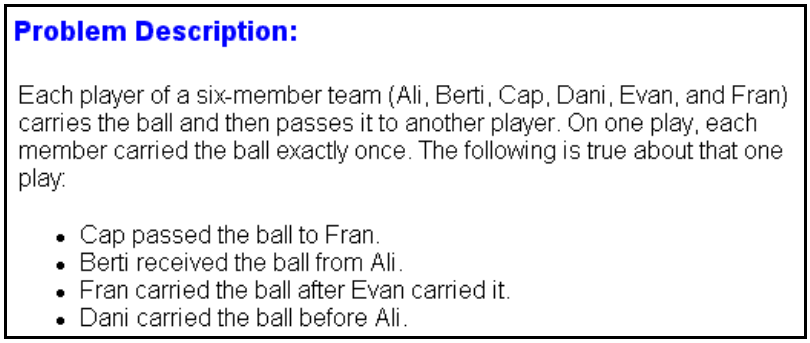
comprehension, reasoning, and test-taking skills. The second test was presented in T vs. GV mode to compare the differences between these two presentation modes.

Participants were evenly and randomly distributed into two treatment groups. Their background information was recorded to ensure a controlled balance in demographic characteristics between groups. Because individual participants' analytical comprehension and reasoning, and task-taking skills varied greatly and such skills could affect their performance in the experiment task, a 15-minute GRE analytical test was conducted before the actual experiment task. All information for the pre-test was visually presented as texts on a Web page for both groups. Subject's performance on this pre-test was used as a covariate in the analysis of the experiment task performed later.

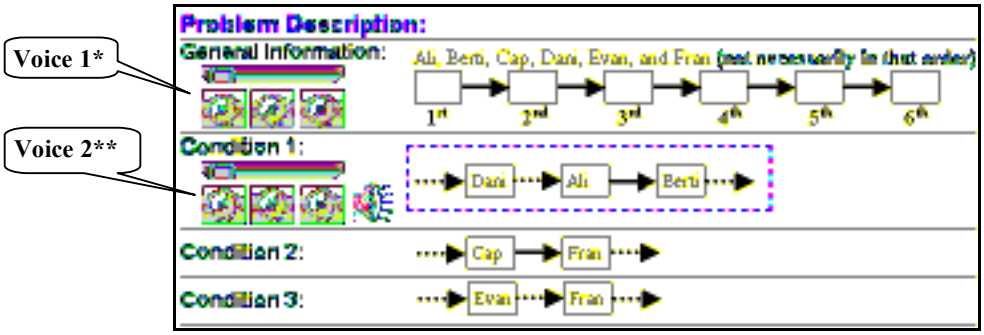
4.3. Independent and dependent variables

The only independent variable was information presentation mode. There were two treatments: Text (T) mode and Graphic + Voice (GV) mode. In the T-mode, all information was visually presented as text on a Web page. In the GV-mode, the original textual information was split into a flowchart-like diagram and speech output. Three faculty members with rich teaching experience were asked to manually convert the GRE analytical tests into a graph + voice presentation according to the proposed method (see figure 3). Only the internal relationships and related entities were converted into graphics. An example of the information presented in the T-mode and GV-mode is shown in figure 5.

T-mode Information Presentation



GV-mode Information Presentation



*Voice 1: "Each player of a six-member team (Ali, Berti, Cap, Dani, Evan, and Fran) carries the ball and then passes it to another player. On one play, each member carried the ball exactly once. The following is true about that one play."

**Voice 2: "Dani carried the ball before Ali, and Berti received the ball from Ali. Similar diagrams are used for condition 2 and 3." [The image icon and the blue dotted-line frame of the diagram indicate voice 2 is currently playing.]

Figure 5. An Example of Display Modes

The two dependent variables were user performance and satisfaction. User performance was measured by the number of correctly answered questions within a 30-minute period. The task started when the first analytical problem was presented on the screen, and ended when time was up. User satisfaction was measured by a satisfaction questionnaire using a 7-point Likert scale, based on technology acceptance model (TAM) (Davis 1989, Koufaris 2002). This satisfaction questionnaire was designed to measure the user's perceived usefulness and ease of use of the two interfaces. In addition, one question was added to measure user overall satisfaction.

4.4. Procedure

Each subject was asked to sign a consent form before participation. During the training session, each subject filled out a background questionnaire and then the experimenter described the tasks included in different groups. A sample problem was used to explain the interface, browsing rules, time limit, graphic notations (for the GV-mode group), and voice control (for the GV-mode group). Subjects were allowed to ask questions and spend as much time as needed during the training session. They were encouraged to answer as many questions as they could during the two analytical tests. They were allowed to browse back and forth within each problem to find or correct their answers. Subjects could click a submit button to move on to the next analytical problem after they finish the current one, but they could not return to the previous problem. For the GV-mode presentation, pre-recorded voice information was automatically played when the Web page was loaded on the screen. Subjects could click the control buttons on the screen to replay voice messages. Subjects were allowed to take breaks before and after the timed tests. Upon completing the two tests, the subject was asked to fill out a satisfaction questionnaire. There was no time limit for this satisfaction survey. Table 5 presents the experiment procedure.

Table 5. Experiment Procedure

	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 in T-mode presentation	Solve problems in GV-mode presentation	Satisfaction questionnaire

The following information was collected during the experiment and saved into a database:

- 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, and
- Subjects' online activities (e.g., manipulating voice messages, changing answers).

5. Results and discussion

5.1. Task Performance

The hypothesis postulates that both user comprehension of information and satisfaction will be improved by the dual-modal presentation. The first dependent variable, task performance, was measured by the number of correctly answered questions in the experiment task. Table 6 presents descriptive statistics of task performances.

Table 6. Descriptive Statistics of Task Performances

	Pre-Test				Task			
	T-Mode		GV-Mode		T-Mode		GV- Mode	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Number of correct answers	5.8	1.36	5.8	2.77	9.9	3.58	19.1	9.42
Accuracy	0.835	0.1024	0.709	0.1455	0.456	0.1253	0.581	0.1918

Considering the individual differences, participants' performance in the pre-test was used as a covariate to adjust the results of their task performance. The normality and constant variance assumptions were verified. A logarithm transformation was applied on task performance and a square root transformation was applied to task accuracy to ensure constant variances. The adjusted task performance was used in the analysis of covariance. Results are shown in table 7. A significant difference was found in the number of correctly answered questions between the T-group and GV-group (T-Mode: mean=9.9, GV-Mode: mean=19.1, $p<0.0001$). As shown in table 6, on average the participants in the GV group correctly answered twice as many questions correctly as their peers in the T group did. Table 6 also provides an additional comparison between the accuracy, defined as the number of correctly answered questions divided by the total number of answered questions, of the performance in the two groups. Accuracy, although not a dependent variable in this experiment, was measured for precaution because if a participant guessed many times during the task, he or she might be able to correctly answer more questions with a lower rate of accuracy. The average probability of guessing the correct answer of one question would be 0.20 (5 choices for each question). The ANCOVA results of accuracy was presented in Table 8. The significant difference in accuracy (T-mode: mean=0.456, GV-mode: mean=0.581, $p < 0.001$) indicates that the greatly improved performance in the GV-group was not the results of random guessing. The accuracy on the GV-mode presentation on average increased by about 30%, as compared to the accuracy on the T-mode presentation. These results confirm that participants in the GV-group did perceive and process information more effectively within the given time.

Table 7. ANCOVA of Task Performance

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	2.4379	2.4379	25.12	<0.0001
Pre-test	1	2.0867	2.0867	21.50	<0.0001

Table 8. ANCOVA of Task Accuracy

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	0.2276	0.2276	13.90	0.0010
Pre-test Accuracy	1	0.1857	0.1857	11.34	0.0025

To further determine the significant difference in task performance between the T and GV groups, we also analyzed the time spent on individual questions. Both the pre-test and the experiment task contain multiple problems, each consists of a few questions. Table 9 shows a summary of the average time spent on each problem and each question based on both problems commonly answered by all participants and all problems answered by an individual participant. Information presented in Table 9 provides more details of user task performance.

Table 9. Average Time Per Problem and Per Question

	Pre-Test				Task			
	T-Mode		GV-Mode		T-Mode		GV- Mode	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Average time / problem (sec.) (Based on problems commonly answered by all participants)	438.9	30.32	393.1	79.04	455.4	95.91	340.2	89.22
Average time / question (sec.) (Based on problems commonly answered by all participants)	39.9	2.76	35.7	7.19	26.8	5.64	20.0	5.25
Average time / problem (sec.) (Based on all problems answered by an individual participant)	430.1	52.80	357.1	95.95	402.1	85.50	305.8	70.27
Average time / question (sec.) (Based on all problems answered by an individual participant)	112.6	37.28	84.3	21.81	88.5	27.3	60.4	19.7

ANCOVA analyses were performed and results are presented in tables 10-13. These results suggest that on average, GV-mode users spent significantly less time on each problem (T-mode: mean=455.4 seconds, GV-mode: mean=340.2 seconds, p=0.0203) and on each question (T-mode: mean=26.8 seconds, GV-mode: mean=20.0 seconds, p=0.0203) than T-mode users did based on problems commonly answered by all participants. The results also suggest that on average, GV-mode users spent significantly less time on each problem (T-mode: mean=402.1 seconds, GV-mode: mean=305.8 seconds, p=0.0383) than T-mode users did based on all problems answered by an individual participant. These findings attest that GV mode users had performed the task more efficiently than T-mode users.

Table 10. ANCOVA of Average Time Spent on Each Problem
(Based on Problems Answered by All Participants)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	37743.95480	37743.95480	6.09	0.0203
Time/problem	1	72745.81648	72745.81648	11.73	0.0020

Table 11. ANCOVA of Average Time Spent on Each Question
(Based on Problems Answered by All Participants)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	130.6019197	130.6019197	6.09	0.0203
Time/question	1	251.7156279	251.7156279	11.73	0.0020

Table 12. ANCOVA of Average Time Spent on Each Problem
(Based on All Problems Answered by an Individual Participant)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	23876.53534	23876.53534	4.74	0.0383
Time/problem (for individual)	1	35537.64981	35537.64981	7.06	0.0131

Table 13. ANCOVA of Average Time Spent on Each Question
(Based on All Problems Answered by an Individual Participant)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	1491.262669	1491.262669	3.70	0.0649
Time/problem (for individual)	1	5031.146408	5031.146408	12.49	0.0015

5.2. User Satisfaction

User satisfaction was measured by a questionnaire derived from the technology acceptance model (TAM) (Davis 1989). According to this model, perceived ease of use and perceived usefulness are the two determinants of user intention to adopt a new technology. If a user feels more satisfied with ease of use and usefulness of a technology, it is more likely that he or she will adopt the technology. Therefore, it is reasonable to use perceived ease of use and perceived usefulness as surrogates for satisfaction. In the satisfaction questionnaire, six questions measuring perceived usefulness and seven questions measuring perceived ease of use were drawn from Davis' (1989) study. One question measuring users' overall satisfaction was added, 'In general, I am satisfied with visual-auditory (or textual) presentation of the problems'.

A factor analysis with varimax rotation was performed to establish convergent and discriminant validity of the two constructs. Four items of low loadings were removed from the original seven questions measuring perceived ease of use (PE). Cronbach's α values were calculated to verify reliability of the instrument. The high Cronbach's α values for perceived ease of use ($\alpha = 0.88$) and perceived usefulness ($\alpha = 0.97$) suggest that the questionnaire is reliable and valid.

After dropping the four items of PE, 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 perceived usefulness, perceived ease of use, and overall satisfaction between the two groups. Overall satisfaction was transformed to satisfy the homogeneity of variance assumption of t-test. Table 14 presents the results of these t-tests and the descriptive statistics of perceived usefulness, perceived ease of use, and overall satisfaction.

Table 14. Descriptive Statistics and T-test Results of Satisfaction

	T-Mode (n=15)		**GV-Mode (n=15)**			
	Mean	**Std.**	**Mean**	**Std.**	**t**	**Pr>|t|**
Perceived Usefulness	21.4	8.89	36.5	4.96	5.74	<0.0001
Perceived Ease of Use	15.8	3.45	19.3	2.05	3.35	0.0023
Overall Satisfaction	3.5	1.81	6.2	0.68	5.49	<0.0001

Significant differences between the GV and T modes were found in perceived usefulness ($t(28)=5.74,\ p<0.0001$), perceived ease of use ($t(28)=3.35,\ p=0.0023$), and overall satisfaction ($t(28)=5.49,\ p<0.0001$). Six items were used in the questionnaire to measure perceived usefulness and three items were used for perceived ease of use. Therefore, the average score of perceived usefulness in the GV-group was 36.5/6≈6.1, and 21.4/6≈3.6 in the T-group. The average score of perceived ease of use in the GV-group was 19.3/3≈6.4, and

15.8/3≈5.3 in the T-group. The average score of overall satisfaction was 6.2 in the GV-group, and 3.5 in the T-group. In the 7-point Likert scaled questionnaire, the neutral score is 4. Overall, GV users expressed much higher satisfaction and perceived usefulness on the proposed dual-modal presentation.

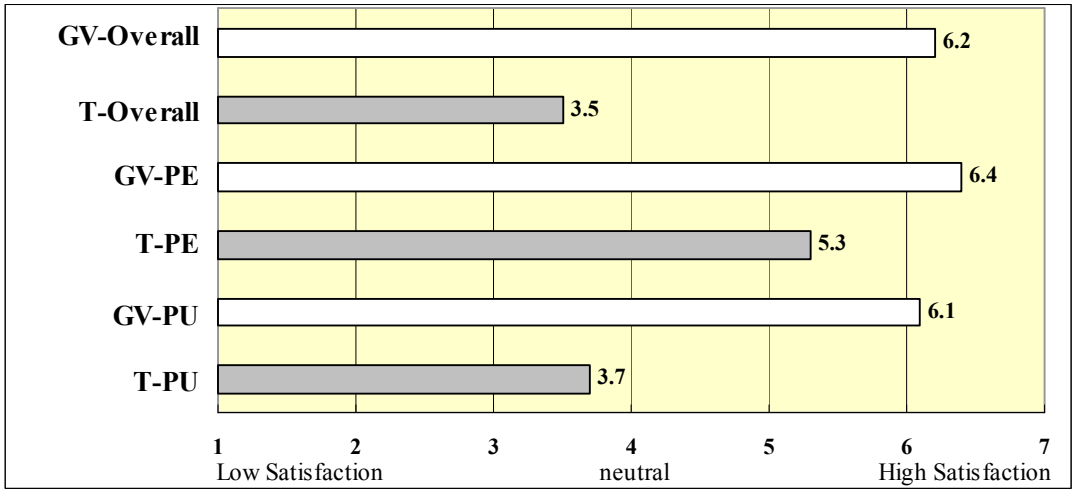


Figure 6. User Satisfactions

5.3. Discussion

According to the results of task performance and user satisfaction discussed in the previous sections, the hypothesis was fully supported by this experiment. The main findings are summarized below:

1. On average participants in the GV group correctly answered about twice as many questions as participants in the T group did.
2. The accuracy on the GV-mode presentation was increased on average by about 30%, as compared to the accuracy on the T-mode presentation.
3. Participants in GV-group spent significantly less time on reading the problem description, on working out each problem, and on answering each question.
4. Participants in the GV-group were significantly more satisfied than their peers in the T-group with their task experience in terms of perceived usefulness, perceived ease of use, and overall satisfaction.
5. Participants' online behavior and their after-task feedback indicated that voice information was helpful and reinforced their understanding of the diagram.

These empirical results are consistent with prior theories in human attention and working memory. As suggested by Wickens' multiple resource pool model (1980 and 1984), 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 working memory model (1986) also predicts that tasks which concurrently use different subsystems (visuo-spatial sketch pad and phonologic loop) in the working memory should have minimum interference. While users are browsing imagery or graphical information, they may be able to receive verbal information from the auditory channel. Furthermore, the results agree with Paivio's dual coding theory (1986) 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. Our study shows that the proposed dual-modal interface is theoretically sound and empirically validated.

6. Conclusions

In this study, we aimed to improve the effectiveness of presenting information using multiple modalities with minimum interference. The proposed dual-modal information presentation was tested through controlled experiments. Findings from this study suggest the following: (1) Users could concurrently integrate perceived visual (diagrams) and auditory (voice messages) input without interference; (2) Highly relevant speech information might facilitate user's understanding of diagrammatic information; (3) The distribution of cognitive workload across modalities might demand less mental effort required in information comprehension, as compared to single-modality presentation; and (4) Users might achieve higher satisfaction due to the alleviated working memory load during information processing.

6.1. Contributions

Findings of this study have profound implications for future research in multimodal interface design. Multimodal interfaces are especially promising to mobile applications due to 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 small screen on which only a limited amount of information can be presented. When the device is used on the move, it makes reading textual information much more difficult. Multimodal interfaces will be able to address these constraints by delivering information through multiple sensory modalities such as visual and auditory channels. The advancement of speech synthesis technology can support information delivery via the auditory channel. Meanwhile, the amount of text could be greatly reduced after being converted into diagrams. Increased readability with a decreased requirement of screen real estate is expected on the graphic visual presentation.

Contribution of this research goes beyond interface design for handheld devices. Results of this study can facilitate the design for generic human-computer interaction and for instructional information presentation. Because visual and auditory perceptual processing are closely linked, perception of disparate information from different modality channels often introduces interference and distraction. Mental integration of the dual-model information also demands a heavy cognitive memory load and leads to degraded performances as often found in computer-based applications with visual-auditory output. Researchers have spent years exploring different aspects of how to efficiently utilize human's 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 is derived from the same textual content. Thus, mental integration of the relevant visual and auditory information does not demand a high cognitive workload.

6.2. Limitations and future research directions

This study has several limitations. First, this experiment focused on the dual-modal presentation of sequential relationship in texts. This might limit the generalizability of the findings to other applications. As suggested by Dubois and Vial (2000), the effectiveness of a multimodal information presentation can be affected by factors such as the presentation mode, the construction of co-referenced materials, and the task characteristics. Second, sample problems taken from the GRE tests were used in the experiments to measure subjects'

analytical comprehension and reasoning skills. However, caution should be exercised when applying findings from this study to different information processing tasks. Third, although its results are consistent with prior cognitive load theories, this study has not attempted to measure cognitive load directly. This study aimed at using these theories to generate practical guidelines for the design and development of effective dual-modal interfaces. The results indicate that the experiment has succeeded in this aim. Nevertheless, there may be plausible alternative explanations of these results. For instance, was user comprehension improved mainly because the clues were integrated into the diagrams? Which component, the voice messages or the graphics, played a more active role in affecting user comprehension in the proposed dual-modal interface?

In the future, we will continue to explore different methods of converting texts into graphics. Advanced techniques will be employed to generate rules of text conversion and speech synthesis. This process should eventually be automated and applicable to a wider range of textual information. Future work may also include subjective measures of cognitive load and the efficiency assessment of individual instructional components. It would be interesting to further investigate how graphics and voice affect user performance and satisfaction separately. Although 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. Users with different occupations and educational levels should be observed in future studies to explore possible different behaviour patterns when they perform comprehension tasks on the proposed dual-modal interface.

References

Archer, N., Head, M., Wollersheim, J. and Yuan, Y., 1996, Investigation of voice and text output modes with abstraction in a computer interface. *Interacting with computers*, 8, 323-245.

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

Baddeley, A. D., 1992. Working memory. *Science*, 225, 556-559.

Baggett, P. and Ehrenfeucht, A., 1983, Encoding and retaining information in the visual and verbals of an educational movie. *Educational communication and technology*, 31, 23-32.

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

Chalfonte, B., Fish, R. and 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. and Lam, J., 2002, Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research,* 3, 187-199.

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.

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

Coull, J. and Tremblay, L. E., 2001, Examining the specificity of practice hypothesis: Is learning modality specific? *Research quarterly for exercise & sport*, 72, 345-354.

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

DeHaemer, M. and Wallace, W., 1992, The effects on decision task performance of computer synthetic voice output. *International journal of man-machine studies*, 36, 65-80.

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

Dubois, M. and 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.

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

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

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

Koufaris, M., 2002, Applying the technology acceptance model and flow theory to online consumer behaviour. *Information systems research*, 13, 205-233.

Logie, R., Gilhooly, K. and 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, 43-64.

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

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

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

Nardi, B. A., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S. and 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. and Gopher, D., 1979, On the economy of the human-processing system. *Psychological review*, 86, 214-255.

Nugent, G., 1982, Pictures, audio, and print: symbolic representation and effect on learning. *Educational Communication and Technology*, 30, 163-174.

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).

Polson, M. C. and Friedman, A., 1988, Task-sharing within and between hemispheres: a multiple-resources approach. *Human factors*, 30, 633-643.

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

Schlosser, R. and Belfiore, P., 1995, The effects of speech output technology in the learning of graphic symbols. *Journal of applied behavior analysis*, 28, 537-549.

Sipior, J. and Garrity, E., 1992, Merging expert systems with multimedia technology. *Database*, 54-49.

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

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

Travers, M., 1989, A visual representation for knowledge structures. In *Proceedings of the second annual ACM conference on hypertext*, pp.147-158.

Tremblay, L. and Proteau, L., 1998, Specificity of practice: The case of powerlifting. *Research quarterly for exercise and sport*, 69, 284-28.

Treviranus, J. and 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 *Attention and Performance VIII*, R. S. Nickerson (Ed.), pp.239-257 (Hillsdale, NJ: Lawrence Erlbaum).

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

Wickens, C. D. and Liu, Y., 1988, Code and modalities in multiple resources: A success and a qualification. *Human factors*, 30, 599-616.

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

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