A Dual-Modal Presentation of Sequential Relationships in Texts

Shuang Xu[a], Xiaowen Fang[b], Jacek Brzezinski[b], Susy S. Chan[b]

[a] *Center for Human Interaction Research, Motorola Labs, Schaumburg, IL 60196, USA*
[b] *School of Computer Science, Telecommunications, and Information Systems, DePaul University ,243 South Wabash Avenue, Chicago, IL 60604, USA*

Abstract

Based on Braddeley's working memory model and research on human attention, this study intends to develop and validate a visual-auditory information presentation that: (1) minimizes the interference in information processing between the visual and auditory channels; and (2) improves the effectiveness of mental integration of information from different modalities. The Baddeley model suggests that imagery spatial information and verbal information can be concurrently held in different subsystems in human's working memory. Accordingly, this research proposes a method to split textual information containing sequential relationships into a "graphic + voice" representation. We hypothesize that this dual-modal presentation will result in superior user comprehension performance and higher satisfaction as compared to pure textual display. An experiment was carried out to test the hypothesis. The independent variable was the presentation mode: textual display vs. visual-auditory presentation. The dependent variables were user performance and satisfaction. Twenty-eight subjects participated in this experiment. The results indicate that the "graphic + voice" presentation significantly improved both users' information processing performance and their satisfaction. Results of this study will benefit the interface design of generic computer systems by alleviating the overabundance of information output in the visual channel. These findings may also help address the usability problems associated with small-screen computers.

Keywords: Multi-modal interfaces, information presentation, human attention, working memory, 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 supported by the independent nature of multi-modal information processing, which assumes that there will be no interference between tasks and thus no degradation in performance [1]. However, research in cognitive psychology shows that visual and auditory perceptual processing is closely linked [2]. Problems related to memory and cognitive workload are found in current applications with voice-based interface [1].

The objective of this research is to develop and validate a dual-modal interface that: (1) minimizes the interference in information processing between visual and auditory channels; and (2) improves the effectiveness of mental integration of information from different modalities. This study focuses on the dual-modal presentation of textual information that describes sequential or chronological events. Findings of this study may benefit the interface design of generic computer systems by alleviating the information overload in the visual channel.

2. Literature review

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 [3]	Attention model	Bottleneck models suggest that only a limited amount of information can be brought from the sensory register to working memory.
Kahneman [4]	Attention model	Resource models view attention as a limited-capacity resource that can be allocated to one or more tasks.
Navon & Gopher [5]; Wickens [6,7]	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.
Wickens, Gordon, & Liu [8]	Multimodality	The amount of shared resources affects how well people can divide their attention between tasks.
Dubois & Vial [9]; Treviranus & Coombs [10]; Coull & Tremblay [11]	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 & Friedman [12]; Wickens & Liu [13]	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 [14] proposes a working memory model that depicts three components: central executive, visuo-spatial sketchpad, and phonological loop (see Figure 1)

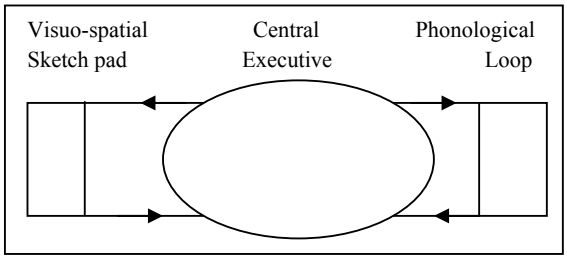

Fig. 1: Baddeley's Working Memory Model (1986)

According to this model, human working memory contains two subsystems for storage: a phonological loop and a 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 [15] show that students' performance was significantly improved when the verbal representation and image representation of a geometry problem were respectively presented in auditory and visual modes.

2.3. Visual and auditory information presentation

Introducing voice into the traditional visual interface presents new challenges. Both the nature and user's 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.

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 characteristics of the task. Findings from multi-modal interface

design studies might not hold true when the content or the nature of the information processing task changes.

Table 2. Research in Visual and Auditory Information Presentation

Author(s)	Findings
Dubois & Vial [9]	The integration of sound, written words, and the image of verbal information results in a light cognitive overload, which improves the effectiveness of learning.
Nardi et al. [16]	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.
Streeter [17]	The major advantages of speech-based interface are universality and mobile accessibility, whereas its disadvantage is the slow delivery rate of voice information.
Treviranus & Coombs [10]	The integration of captioning, video description, and other access tools for interactive information exploration makes the learning environment more flexible and engaging for students.

Table 3. Research in Graphical Representation of Texts

Author(s)	Findings
Mayer [18]; Mayer & Gallini [19]; Mayer & Anderson [20]	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 relationship among them.
Paivio [21,22]	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 & Van Zandt [23]; Johnson-Laird [24,25]	Schemas represent the structure of a person's interest and knowledge, which enables a person to develop the expectancy about what will occur.

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

Denis' [26] finding suggests that imagery of the sequence of events may help users form schemas by reducing the cognitive demand for converting textual information into effective schemas. Knowledge-based information systems are often represented as diagrams of interrelated units connected by lines. Travers' study [27] indicates that when the diagrammatic representation of knowledge structures matches users' 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.

3. Proposed dual-modal information presentation

A new dual-modal presentation was proposed in Figure 2. In this dual-modal presentation, sequential relationships contained in texts will be extracted and presented in diagrams. The remaining textual information will be delivered through the auditory channel. According to Baddeley's working memory model, the verbally represented information will be perceived via the auditory sensory register, then sent to and processed in the phonological loop. Concurrently, 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 can thus be retrieved from the phonological loop and visuo-spatial sketchpad simultaneously, and integrated by the central executive for comprehension.

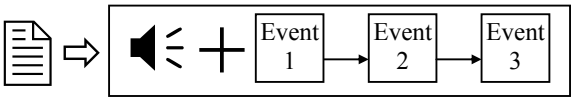

Fig. 2 : Proposed Dual-modal Presentation

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

A reduced cognitive workload is expected because the proposed dual-modal presentation concurrently utilizes the two subsystems in working memory to process the same amount of information. In the proposed presentation, 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.

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. A Web site that presented the GRE analytical tests for the experiment was built. Subjects used this site to perform the task of GRE tests.

4.1. Subjects

Twenty-eight (28) participants were recruited from a Midwest university in the United States. The sample was composed of undergraduate/graduate students, staff, faculty members, and alumni. These participants represented a wide range of background in terms of age, ethnicity, computer experience level, and Internet usage. With a controlled balance in gender and language (see Table 4), participants were randomly assigned to one of the two treatment groups: textual display (T-mode) group and "graphics + voice" display (GV-mode) group.

Table 4. Distribution of Participants

Group	Female	Male	Native English speakers	Non Native English speakers
T-mode	7	7	10	4
GV-mode	7	7	10	4

4.2. Experiment design and the tasks

The experiment design was a 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 for comparing the differences of these two presentation modes.

Because individual participants' analytical comprehension and reasoning skills may vary greatly and such skills could affect their performance in the experiment, an independent GRE analytical test was conducted as a pre-test to estimate a participant's skills before the actual experiment task was performed. In this 15-minute pre-test, all information was visually presented as texts on a Web page for both groups. The estimate of analytical comprehension and reasoning skills or possibly other test-taking skills from the pre-test 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 texts 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 sequential relationships and related entities were converted in graphics.

The two dependent variables were user performance and satisfaction. User's 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) [28,29], this questionnaire was designed to measure user's perceived usefulness and ease of use of the two interfaces. In addition, one question was added to measure user's general satisfaction.

4.4. Procedure

Table 5 presents the experiment procedure. Subjects were encouraged to answer as many questions as they could during the two analytical tests. For the GV-mode presentation, pre-recorded

voice information was automatically played when the Web page is loaded on the screen. Subjects could use controls on the screen to replay voice messages. Subjects were allowed to take breaks before and after the timed tests. Upon completion of these two tests, the subject was asked to fill out a satisfaction questionnaire.

Table 5. Experiment Procedure

	Pre-Test (15 minutes)	Task (30 minutes)	Satisfaction Survey
T-mode Group	In T-mode presentation	In T-mode presentation	Satisfaction questionnaire
GV-mode Group	In T-mode presentation	In GV-mode presentation	Satisfaction questionnaire

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	1.37	5.5	2.47	10.1	3.63	18.1	8.79
Accuracy	0.686	0.1547	0.488	0.1842	0.447	0.1248	0.557	0.1738

5. Results and discussion

5.1. Performance

Table 6 presents descriptive statistics of task performances. Considering the individual difference, participants' performance in the pre-test was used as a covariate to adjust the results of their performance in the Task. The normality and constant variance assumptions were verified. A logarithm transformation was performed to task performance and a square root transformation was applied to task accuracy to ensure constant variances.

The adjusted performance of Task was used in the analysis of covariance and results are shown in Table 7. A significant difference was found in the number of correctly answered questions between the T group and the GV group (T Mode: mean=10.1, GV Mode: mean=18.1, $p<0.0001$). As shown in Table 6, on average the participants in the GV group correctly answered about 1.8 times of the questions that were correctly answered in the T group. 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 is listed in Table 8. Accuracy was not a dependent variable in this experiment. Accuracy was measured for precaution because if a participant guessed a lot during the performance, he/she might be able to correctly answer more questions with a much lower accuracy. The average probability of hitting the correct answer would be 0.20 (5 choices for each question). The significant difference in accuracy (T mode: mean=0.447, GV mode: mean=0.557, $p = 0.001$) indicates that the greatly improved performance in the GV group was not the results of random guessing. On average, the accuracy on the GV-mode presentation was increased about 25%, as compared to the accuracy on the T-mode presentation. These results strongly indicate 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	<.0001
Pre-test	1	2.0867	2.0867	21.50	<.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	1	0.1857	0.1857	11.34	0.0025

5.2. Satisfaction

User satisfaction was measured by a questionnaire derived from the technology acceptance model (TAM) [28]. In the satisfaction questionnaire, six items measure perceived usefulness (PU), three items measure perceived ease of use (PEOU), and one item measures user's general satisfaction.

Table 9 presents the results of these t-tests and the descriptive statistics of perceived usefulness, perceived ease of use, and overall satisfaction.

Table 9. Descriptive Statistics and T-test Results of Satisfaction

	T-Mode (n=14)		GV-Mode (n=14)		t	Pr>|t|
	Mean	Std.	Mean	Std.		
PU	22.2	8.62	36.2	5.04	5.24	<.0001
PEOU	15.9	3.57	19.3	2.13	3.09	0.0048
Overall	3.6	1.83	6.1	0.66	5.62	<.0001

Significant differences between the GV and the T modes were found in perceived usefulness (*t(26)=5.24, p<0.0001*), perceived ease of use (*t(26)=3.09, p=.0048*), and general satisfaction (*t(26)=5.62, p<0.0001*). The average score of perceived usefulness in the GV group was 36.2/6≈6.0, and 22.2/6=3.7 in the T group. The average score of perceived ease of use in the GV group was 19.3/3≈6.4, and 15.9/3≈5.3 in the T group. The average scores of overall satisfaction was 6.1 in the GV group, and 3.6 in the T group. In the 7-point Likert scaled questionnaire, the neutral score is 4. These results strongly indicate that participants in the GV group were generally satisfied with the information presentation.

6. Conclusions

In this study, we aimed to improve the effectiveness of presenting information using multiple modalities with minimum interference. A dual-modal information presentation was proposed and validated through a controlled experiment. The results agree with prior research findings. Based on the multiple-resource human attention model [6,7], two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. According to Baddeley's working memory model [14], tasks using different subsystems in the working memory should not interfere with each other. The results of our experiment also conform to Mousavi's study [15], where students' performance was significantly improved when the verbal representation and image representation of a geometry problem were respectively presented in auditory and visual mode.

Findings from this study suggest the following: (1) Users might be able to concurrently integrate perceived visual (diagrams) and auditory (voice messages) input without interference; (2) Highly relevant speech information might facilitate the user's understanding of diagrammatic information; (3) The distribution of cognitive workload across modalities might reduce the demand for mental efforts required in information comprehension, as compared to single-modality presentation; and (4) Users might have higher satisfaction due to the alleviated working memory load during information processing.

References

[1] 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.

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

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

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

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

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

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

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

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

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

[11] 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.

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

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

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

The complete list of references is available at http://condor.depaul.edu/~xfang/iea2006/references.htm