Investigation of Dual-Modal Information Presentation

Shuang Xu, Kimberly Watson, Xiaowen Fang, Susy Chan, Jacek Brzezinski
DePaul University
School of Computer Science
Chicago, IL
shuangxu@yahoo.com, kwatson@cs.depaul.edu, xfang@cs.depaul.edu, schan@cs.depaul.edu, Jbrzezinski@cs.depaul.edu

Abstract

To design computing interfaces enabling users to best receive information, the effects of multi-modal information presentation on user task performance and acceptance must first be ascertained. Based upon human attention models and prior research findings, this paper investigates dual-modal information presentation designed to test users' ability to receive information via auditory channel in addition to normal visual mode. By examining users' task performance and the extent to which users will accept dual-modal information presentation, this investigation seeks to establish a theoretical foundation for the design of multi-modal interfaces to allow users to access rich multimedia information space in the future.

1 Introduction

The majority of displays encountered in human-machine systems are single-modal - either visual or auditory. However, multi-modal interfaces with both visual and auditory output are considered by many to have a very bright future [6] [7] [9]. Multi-modal interfaces are especially promising for applications lacking sufficient screen space for comprehensive visual display and for applications requiring access by users under mobile conditions.

The objective of this investigation is to examine the effectiveness and feasibility of dual-modal information presentation that provides relevant data through the auditory channel in addition to the regular visual display. The study will likely establish a theoretical foundation for the design of multi-modal interfaces by determining the extent to which users can effectively process information from multiple sources. These findings will help researchers and developers to maximize the effectiveness of information presentation to a wide variety of users through the incorporation of visual and auditory cues in application programming.

2 Background literatures

2.1 Human attention

The extent to which humans are able to pay attention has long been of interest to researchers. Proctor and Van Zandt [11] distinguish human attention in three aspects: selective attention, concerning human ability to focus on certain sources of information and ignore others; divided attention, involving human ability to divide one's attention among multiple tasks; and the

amount of mental effort required to perform a task. Several theoretical models pertaining to attention have been proposed. Bottleneck models specify a particular stage in the information-processing sequence where an individual's ability to attend to incoming information becomes limited. In contrast, resource models view attention as a limited-capacity resource that can be allocated to one or more tasks, rather than as a fixed bottleneck. Among various attention models, multiple-resource models propose that there is no single attention resource. Rather, several distinct subsystems each have their own limited pool of resources. Wickens [16] [17] proposes a three-dimensional system of resource utilization consisting of distinct stages of processing (encoding, central processing, and responding), codes (verbal and spatial), and input (visual and auditory) / output (manual and vocal) modalities. The model assumes that two tasks can be performed together more efficiently to the extent that they require separate pools of resources.

2.2 Visual and auditory interfaces

Several studies have investigated the effects of using text, voice, or text combined with voice output modes on user task performance. In comparing speech with text, Streeter [13] points out that the major advantage of using speech in an interface is its universality, because virtually everyone understands spoken language and users can be mobile while listening to a speaker. One notable disadvantage is that voice delivery of information occurs at less than half the rate at which text can be read. In a study of the effects of media (pictures, audio, and print) on student learning, Nugent [10] has found that student learning could be improved by incorporating additional channels such as pictures and audio when the same content was presented in three channels. When different information was presented in the visual and audio modes, however, student learning was not improved by the addition of new channels even though the presence of visual cues did not interfere with processing the audio information and vice-versa. A similar study conducted by Baggett and Ehrenfeucht [2] suggests that there is no competition for resources when related information is presented simultaneously in visual and auditory channels.

Sipior and Garrity [12] indicate that information presentation with a mix of audio and visual accompaniments improve receptiveness attributes such as perception, attention, comprehension, and retention. DeHaemer and Wallace [5] suggest that the visual and audio modes of receiving information appear to be mutually non-interfering and may enhance performance for certain tasks. They observed the effect of voice output on computer-supported decision-making where voice instructions were used to solve a visual decision problem. Furthermore, they found an interaction effect between user decision style and the use of a synthetic computer voice. Comparing voice and text annotation in co-authored documents in terms of interactivity and expressiveness, Chalfonte, Fish and Kraut [3] have found that voice presentation was preferred for addressing higher level issues related to the suggestion of document modifications. However, text presentation was preferred for more detailed and lower level comments. Archer, Head, Wollersheim and Yuan [1] have designed an interface to study user preferences and the effectiveness of output modes. Their findings show that adding text to voice output improves the perceived acceptability of voice, but adding voice to text does not alter the perceived acceptability of text. When the same information was presented in both modes, the text mode was most efficient in performing information searches, followed by the voice mode, and then the text plus voice mode.

3 Proposed dual-modal information presentation

Based on the prior research findings, a dual-modal information presentation, "Visual + Auditory Information" is proposed. In this presentation format, a regular document is displayed in normal visual mode while additional information about this document is presented as voice output. The multiple-resource human attention model proposed by Wickens [16] [17] suggests that two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. Therefore, users may be able to receive brief relevant information from the auditory channel while they are retrieving information visually. Research concerning visual and auditory interfaces imply that voice output that is relevant to the user tasks will not interfere with user attention to visual displays. Accordingly, the following hypotheses will be used to test the effectiveness of this dual-modal information presentation.

3.1 Hypothesis 1

Additional auditory information presented during a web browsing process can be perceived by users and will not negatively impact users' performance on browsing tasks. A simple T-test comparing "Regular Visual Display" and "Visual + Auditory Cues (irrelevant to text-based questions)" will be used to test this hypothesis.

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

3.2 Hypothesis 2

Helpful information presented in the auditory channel will improve users' performance on web browsing tasks. A simple T-test comparing "Regular Visual Display" and "Visual + Auditory Cues (relevant to text-based questions)" will be used to test this hypothesis.

If the average number of correctly answered questions in "Visual + Auditory Cues (relevant to text-based questions)" group is significantly greater than that of "Regular Visual Display" group, then hypothesis 2 is supported.

4 Method

4.1 Experiment design

An experiment is under way to test the hypotheses above. A web site containing generic curriculum information was developed for this experiment. This web site has a simple 3-level navigation hierarchy as follows:

Home (index)
- |__Biomedical Engineering (description of degree requirements)
- | |__Prerequisite Phase (course information in prerequisite phase)
- | |__Foundational Phase (course information in foundational phase)

| |__Advanced Phase (course information in advanced phase)
|__Computer Science (description of degree requirements)
| |__Prerequisite Phase (course information in prerequisite phase)
| |__Foundational Phase (course information in foundational phase)
| |__Advanced Phase (course information in advanced phase)
|__Economics (description of degree requirements)
|__Prerequisite Phase (course information in prerequisite phase)
|__Foundational Phase (course information in foundational phase)
|__Advanced Phase (course information in advanced phase)

Additional curriculum information has been designed and pre-recorded for auditory presentation, and consists of two types; vocal cues intended to assist the user to answer text-based questions about the web site information, and vocal cues presenting generic information to the user that do not provide assistance in answering the text-based questions. The user's task is to browse the web site and listen to the auditory presentation (if any), find information relevant to pre-defined task questions based upon both the text and auditory output, and answer the task questions. Participants perform tasks on a personal computer (PC). A set of questions about the web site and the additional information delivered in auditory mode have been designed to measure the user's information retrieval performance. These pre-defined questions vary randomly from easy ones with answer cues provided in either visual or auditory form to difficult ones with answers to be obtained solely from visual or auditory information.

4.1.1 Independent and dependent variables

The information presentation mode is the only independent variable examined in this experiment. Ninety participants are being randomly assigned into three groups (thirty in each), with different information presentation treatments as follows:

Group A - Regular Visual Display (textual information only)
Group B - Visual + Auditory Cues (relevant to text-based questions)
Group C - Visual + Auditory Cues (irrelevant to text-based questions)

Group A simply views the experiment web site using regular visual display and is asked questions pertaining to the information contained therein. They are able to browse the web site freely using typical navigation links and back/forward buttons in order to find the information necessary to answer their task questions.

Group B has the same browsing capability as Group A, but also receives auditory cues that may direct them to the appropriate page of the website where an answer to a particular question is found, or that may help them to narrow down the choices for the answer based on additional information provided in the voice cue.

Group C is likewise free to browse the website, but the auditory cues they receive relate only to the section of the website they may be browsing at the time. The auditory cues do not assist them in finding the answer to a text-based question; rather, they are presented with information about items such as extracurricular events, course scheduling, or visiting faculty. Furthermore, Group C participants answer additional questions pertaining to the auditory cues they receive interspersed with the task questions posed to Groups A and B.

There are three dependent variables for Hypothesis 1:

- The number of correctly answered questions related to the text-based web site. This variable is used to measure users' performance on browsing tasks.
- The number of correctly answered questions related to the voice cues. This variable is used to measure users' perception of the additional auditory information in Group C.
- Satisfaction. This variable is used to measure users' satisfaction with the display mode.

There are two dependent variables for Hypothesis 2:

- The number of correctly answered questions related to the text-based web site. This variable is used to measure users' performance on browsing tasks.
- Satisfaction. This variable is used to measure users' satisfaction with the enhanced display mode.

	Task Performance		**User Satisfaction**
	Dependent Variable 1	**Dependent Variable 2**	**Dependent Variable 3**
Hypothesis 1	Number of correctly answered questions related to the text-based web site.	Number of correctly answered questions related to the voice cues.	Satisfaction.
Hypothesis 2	Number of correctly answered questions related to the text-based web site.		Satisfaction.

Table 1. Dependent variables

Task performance is measured by the number of correct answers from 25 questions pertaining to the text-based web site (for all three groups) and the number of correctly answered questions from 11 relating to the additional auditory information (for Group C only). User satisfaction is measured by questionnaires utilizing 7-point Likert scales [14] [15]. Groups A and B are asked 12 post-experiment survey questions relating to the perceived usefulness, perceived ease of use and their perceived control of the text-based website with and without relevant voice cues respectively, and Group C is asked 15 post-experiment survey questions pertaining to the perceived usefulness, perceived ease of use and their perceived control of the text-based website with extraneous voice cues [4] [8].

4.1.2 Implementation

For all three experiment groups, a controlled Internet browser window and a task question window are displayed simultaneously on the screen. Participants may switch between the two windows freely, but cannot move or resize either of them. Task questions are presented in the task question window sequentially (i.e., one question at a time in predetermined order) and allow the user to select an answer from four choices, only one of which is correct. The total time allotted for the entire experiment not including the pre- and post-experiment questionnaire and survey is 25 minutes.

Group A participants have the full 25 minutes to attempt to answer 25 task questions correctly based solely upon the information presented in the text-based website. They may browse at leisure, although the Microsoft Windows system clock is displayed in the lower right-

hand corner of the screen to remind them that they have a fixed time limit to attempt to complete the questions. If participants struggle for two minutes and still cannot answer a task question, a "Skip" button automatically appears to allow them to move to the next question without specifying an answer for the current question.

In the presentation for Group B, in addition to the text-based visual display described above, extra information is presented through the auditory channel. Voice cues designed to assist users are played at a random interval of between 6 and 15 seconds after the question first displays (see Figure 1) and are programmed to play in a fixed order that appears random to users in order to minimize the expectation of receiving the voice cues for every question. Users answer the same 25 task questions in the same order as the Group A participants, and have the same browsing capabilities and the opportunity to "Skip" a question after two minutes. No additional questions are presented to Group B relating to the voice cues they hear since these cues relate directly to the text-based task questions.

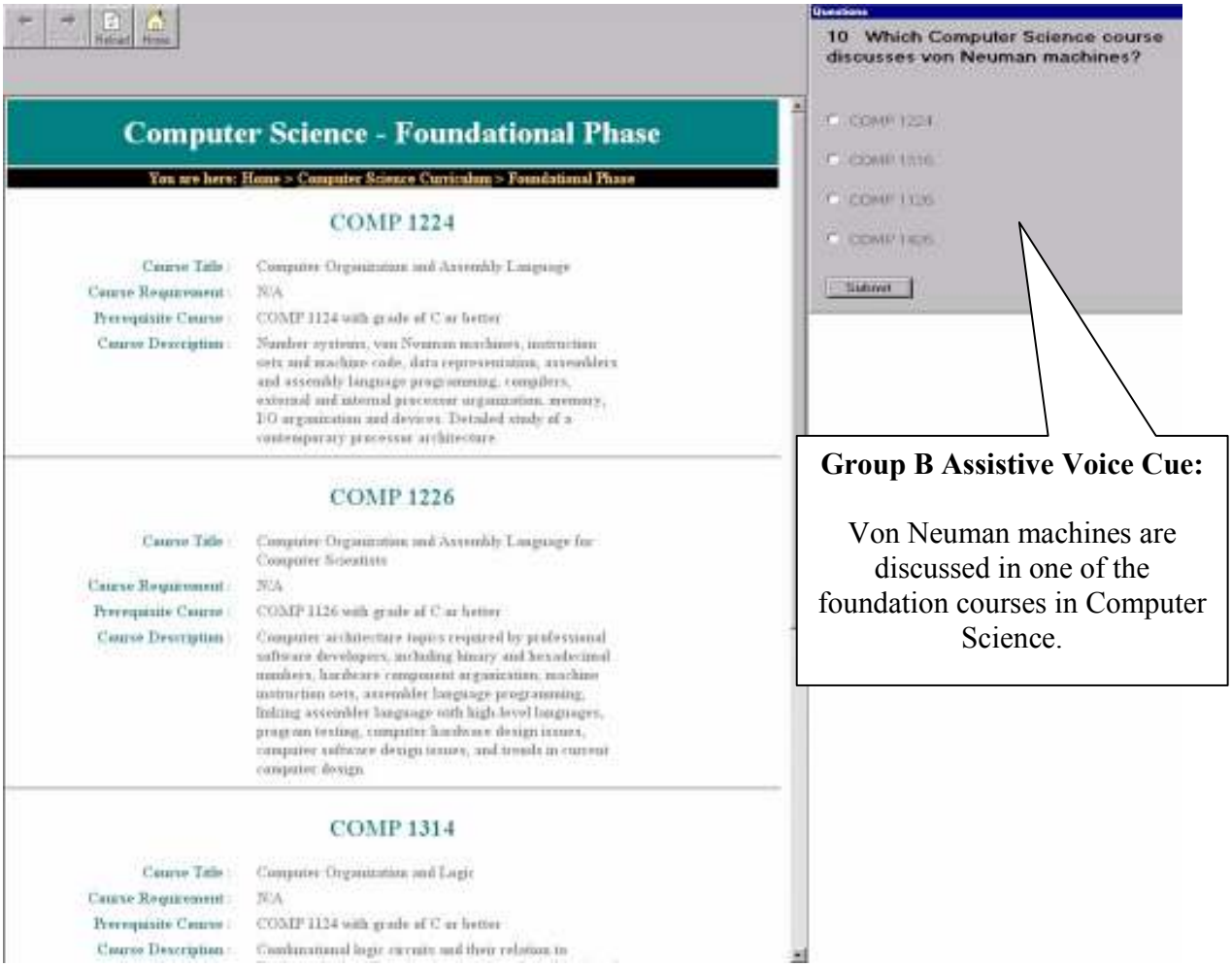

Figure 1. Screen shot of the interface

To discover whether there are differences in information perception between voice cues designed to assist users with their task versus the comprehension of information not directly

related to ongoing tasks, participants in Group C are presented with voice cues that do not assist them to correctly answer the text-based task questions. A Group C voice cue provides information not related to the textual date, and may inform the user about an extracurricular event related to the department the user is currently browsing or about departmental advising hours, for example. These voice cues are played at a random interval of between 10 and 15 seconds after the text-based question displays. To eliminate participants' expectation of voice cues, the interval between the times when the text question displays and when the voice cue plays varies according to the above timing sequences. Although the questions that are followed by voice cues are pre-defined, the sequence appears random to the user. To alleviate synchronization concerns, if participants answer a text question before the voice cue relating to that question plays and he or she moves to the next question, this voice cue is skipped and the experiment continues. By providing the appearance of random ordering and intermixing of voice and text cues to the user, the intent is to focus the user on the experiment since he or she presumably attempts to pay close attention to the auditory channel either for help with text questions (Group B) or to successfully answer a question related to the auditory cue (Group C).

4.2 Subjects

Participants are primarily recruited from the DePaul University student body, which hosts a variety of different age groups, ethnicities, computer experience levels, and knowledge backgrounds. A pre-experiment questionnaire is used to record participants' native language, experience with Internet browsing and voice-based interface, etc. Recognizing that subjects' listening comprehension ability and previous experience with the Internet and with voice interfaces might affect the validity of this experiment, participants are distributed into the three groups to ensure a controlled balance in demographic characteristics.

4.3 Procedure

A pilot study is being conducted for each test group to ensure the proper functioning of the questionnaire, experiment and survey. Preliminary data has been recorded and is being analyzed to make certain that the experiment mechanism and testing procedure are providing accurate and relevant data.

Each participant is asked to sign a consent form before participating in the experiment. Upon completion of a pre-experiment questionnaire, each participant receives instructions for using the experiment browser and performing the requisite tasks. These instructions are presented via computer and in hard copy format. The participant then starts a training session in which he or she browses a sample web site with visual (and auditory) information similar to the version of the experiment the user will encounter. The participant must pass a post-training test consisting of 3 questions for Group A or 5 questions for Groups B and C regarding both the training tasks and the experiment software.

Upon successful completion of the training test, the participant begins the experiment tasks. Considering that fatigue factor might affect participants' understanding capability and listening comprehension, the total duration of the experiment is limited to 25 minutes. According to the best performance in the pilot testing, very few, if any, participants are able to finish all the questions in less than 25 minutes. All participants are expected to correctly answer as many questions as they can in the given time or less. Subjects are informed of the time limit prior to

starting the experiment in order to focus their attention on the tasks at hand, and the program is programmed to automatically terminate at the 25 minute mark. Following the completion of the experiment tasks or the expiration of the experiment time allowed, the participant is asked to complete a survey about his or her perceptions of the ease of use, usefulness, and the perceived control they enjoyed over the information display mechanism. Participants in Groups B and C are then debriefed to provide additional information about their experience in processing and using the voice and text information. The debriefing sessions are recorded and will be transcribed at a later time for data analysis.

5 Next step

The focus of this research is to work toward the establishment of a theoretical foundation and guidelines for the design of multi-modal interfaces that will allow users to receive and comprehend information stemming from both visual and auditory channels. The hypotheses described herein are currently being investigated using the experiment discussed in this paper, and some initial data analysis on the pilot study results is ongoing. Once the pilot study is complete, testing for this experiment will take place over the next two to three months, and data analysis will be conducted afterwards to test the hypotheses.

References

[1] Archer, N., Head, M., Wollersheim, J., & Yuan, Y. (1996). Investigation of voice and text output modes with abstraction in a computer interface. Interacting with Computers, 8(4), 323-345.

[2] Baggett, P. & Ehrenfeucht, A. (1983). Encoding and retaining information in the visuals and verbals of an educational movie. Educational Communication and Technology Journal, 31 (1), 23-32.

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

[4] Davis, Fred D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, September 1989, 319-340.

[5] DeHaemer, M. & Wallace, W. (1992). The effects on decision task performance of computer synthetic voice output. International Journal of Man-Machine Studies, 36, 65-80.

[6] Delogu, C., Paoloni, A., & Pocci, P. (1991). New directions in the evaluation of voice input/output systems. IEEE Journal on Selected Areas in Communications, 9(4), 566-573.

[7] Economist Technology Quarterly (2002). The power of voice. December, 2002, 25-26.

[8] Koufaris, Marios. (2002). Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior. Information Systems Research, Vol. 13, No. 2, 205-223.

[9] Lee, K., Hauptmann, A., & Rudnicky, A. (1990). The spoken word. Byte, July 1990, 225-232.

[10] Nugent, G. (1982). Pictures, audio, and print: symbolic representation and effect on learning. Educational Communication and Technology, 30(3), 163-174.

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

[12] Sipior, J. & Garrity, E. (1992). Merging expert systems with multimedia technology. Data Base, 23(1), 45-49.

[13] Streeter, L. (1988). Applying speech synthesis to user interfaces. In Helander, M. (ed.) Handbook of Human-Computer Interaction (pp. 321-343). New York, NY: Elsevier Science Pub.

[14] Venkatesh, V. (2000). Determinants Of Perceived Ease Of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. Information Systems Research, Vol. 11, No. 4, pp. 342-365.

[15] Venkatesh, V., and F.D. Davis. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science, Vol. 46, No. 2, 186-204.

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

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