A Dual-Modal Information Presentation for Mobile Applications

Xiaowen Fang [a], Jacek Brzezinski [a] , Susy Chan [a], Kimberly Watson [a] , and Shuang Xu [a]

[a] School of Computer Science, Telecommunications, and Information Systems, DePaul University, USA.

Abstract. This research investigated the effectiveness and feasibility of a dual-modal information presentation, "Visual + Auditory (Helpful Info)". Based on the multiple-resource human attention model (Wickens, 1980, 1984), this study hypothesizes that voice hints presented during a web browsing process can be perceived by users and will be used to improve users' performance on the browsing task. A controlled experiment was conducted to test the hypothesis. The hypothesis was partially supported. The findings suggest that users might be able to perceive auditory information while visually browsing textual information.

Keywords. Dual-modal interface, visual, auditory, information presentation.

1. Introduction

The convergence of mobile Internet and wireless communication technology has promised users anytime, anywhere access of information for work and personal communication. However, many technology constraints hinder such access through a mobile device. Two important constraints are the device's small screen size and its mobile use. Compared to desktop or laptop computers, mobile devices typically have a very small screen on which only a very limited amount of information can be presented. When the device is used on the move, it makes reading textual information much more difficult (Chan, Fang, & Brzezinski et al, 2002).

Multimodal interfaces are considered by many to have a very bright future. It has been predicted that in the long run more businesses will use voice applications to supplement or complement text-based e-commerce applications (Economist Technology Quarterly, 2002). Several factors drive this trend: (1) A voice-based interface may increase the potential customer base because there are still more telephones than Web-enabled devices; (2) The screen size is limited on wireless devices; and (3) Customers use their telephones more frequently because they prefer to communicate using natural language. Previous research has examined the effects of presenting information in both auditory and visual modes. In general, the major advantage of using speech in an interface is its universality. The main disadvantage is that human beings can only process voice output in a relatively slow speed (Streeter, 1988). Auditory presentation should be used if the task is performed in continuous motion (Proctor & Van Zandt, 1994). Based on Wickens' (1980, 1984) multiple-resource human attention model, two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. In other words, humans could accept information without interference from two completely different channels: visual and auditory, at the same time. Therefore, if voice output is integrated with visual presentation on a mobile device, problems related to limited screen space and mobility could be addressed by a multimodal interface. Modality integration will require the split of retrieved and summarized information into visual and auditory channels.

The objective of this project is to investigate the effectiveness and feasibility of a dual-model information presentation that outputs relevant information through auditory channel in addition to regular visual display. Findings from this research could contribute to the establishment of guidelines for the design of dual-modal interfaces.

2. Background Literature

2.1. Human Attention

The topic of attention has long been of interest to researchers. Proctor and Van Zandt (1994) distinguish human attention in three aspects: selective attention that concerns human ability to focus on certain sources of information and ignore others; divided attention that involves human ability to divide attention among multiple tasks; and the amount of mental effort required to perform a task. Several models of attention have been proposed. Bottleneck models specify a particular stage in the information-processing sequence at which the amount of information to which humans can attend is 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 (1980, 1984) proposes a three-dimensional system of resources consisting of distinct stages of processing (encoding, central processing, and responding), codes (verbal and spatial), and input (visual and auditory), plus 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

The majority of displays encountered in human-machine systems are either visual or auditory. In general, because spatial discriminations can be made most accurately with vision, spatial information is best conveyed through visual displays (Proctor & Van Zandt, 1994). Likewise, temporal information is best conveyed through auditory displays because temporal organization is a primary attribute of auditory perception.

Several studies have investigated the effects of using text, voice, or text combined with voice output modes. In comparing speech with text, Streeter (1988) points out that the major advantage of using speech in an interface is its universality, because everyone understands spoken language and users can be mobile. One notable disadvantage is that voice delivers information at less than half the rate that text can be read. In a study of the effects of media (pictures, audio, and print) on student learning, Nugent (1982) 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 but the presence of visual did not interfere with processing the audio and vice versa. A similar study conducted by Baggett and Ehrenfeucht (1983) suggests that there is no competition for resources when related information is presented simultaneously in visual and auditory channels.

Sipior and Garrity (1992) indicate that presentation with a mix of audio and visual accompaniments improve receptiveness attributes such as perception, attention, comprehension, and retention. DeHaemer and Wallace (1992) suggest that the visual and audio modes of receiving information appear to be 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, and found an interaction effect between user decision style and the use of computer synthetic voice. In comparing voice and text annotation in co-authored documents in terms of interactivity and expressiveness, Chalfonte, Fish and Kraut (1991) found that voice was preferred for addressing higher level issues in suggesting document modifications, but text was preferred for more detailed and lower level comments. Archer, Head, Wollersheim and Yuan (1996) 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, text mode was most efficient in performing information search, followed by voice mode, and text plus voice mode.

Speech output has also been used widely to assist sight-impaired users. Schlosser, Belfiore, and Nigam (1995) point out that the presentation of additional auditory stimuli in the form of synthetic speech is effective in assisting individuals with mental retardation to learn associations of graphic symbols with spoken words. Gorenflo and Gorenflo (1994) show that attitudes toward the augmented communicator are more favorable in terms of evaluation and potential interaction when the synthetic voice is easier to listen to. Krell and Cubranic (1996) developed a special browser “V-Lynx” with voice output for sight-impaired users. The “V-Lynx” browser reads the first sentence in a paragraph for quick scanning of the document, conveys the document structure (headings, emphasized text, lists, and hyperlinks), and allows for easy navigation while inside and between documents.

Oviatt and Olsen (1994) have studied how people use computers when allowed to use multimodal interfaces for data entry, retrieval, and navigation. The most important factor in predicting the use of multimodal interface is contrastive functionality. For example, users utilize one modality for data input but another for input corrections. A project conducted by Cohen (1992) involved the integration of keyboard, screen, pointing device, and auditory channel. Completion of a task required the utilization of at least two channels such as auditory and keyboard.

3. Proposed Dual-Modal Information Presentation

Based on the prior research findings, a dual-modal information presentation, “Visual + Auditory (Helpful Info)”, was proposed. In this presentation, a regular document is displayed in the normal visual mode while additional information helping understand this document is presented as voice output. The multiple-resource human attention model proposed by Wickens (1980 and 1984) suggests that two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. Hence, users might be able to receive brief relevant information from the auditory channel while they are retrieving information visually. The following hypothesis was used to test the effectiveness of this dual-modal information presentation.

Hypothesis: The “Visual + Auditory (Helpful Info)” presentation will allow users to receive additional helpful information and thus improve the effectiveness of browsing textual information.

4. Method

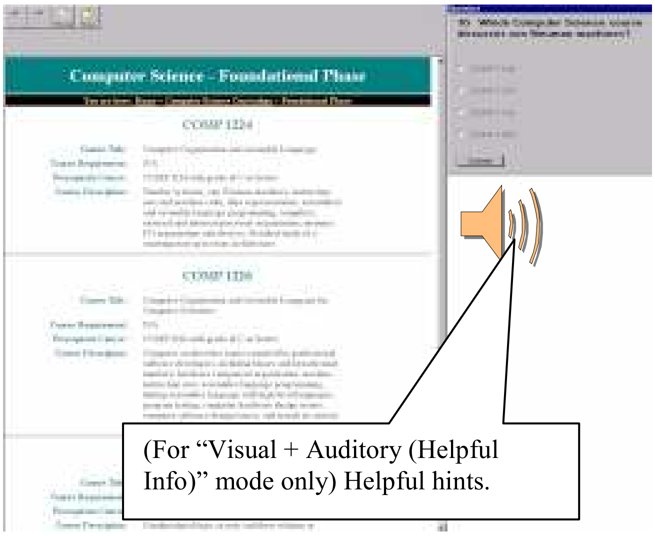
A web site containing generic curriculum information was developed for this experiment. Additional curriculum information was designed and pre-recorded for auditory presentation. The user’s task was to browse the web site and listen to the auditory presentation (if any), find information relevant to pre-defined task questions based upon the textual web site and answer the task questions. Participants performed tasks on a personal computer (PC). A set of questions about the web site were designed to measure the user’s information retrieval performance.

The only independent variable was the information presentation mode. There were two modes: “Regular Visual Display” and “Visual + Auditory (Helpful Info)”. In the “Regular Visual Display” mode, all information was presented visually with no auditory cue. The task for participants was to browse the textual information contained in the Web site and answer questions based on the texts. In the “Visual + Auditory (Helpful Info)” mode, information helping participants to understand the textual information was randomly presented through auditory channel while participants were browsing the textual content of the Web site. In another word, in addition to the primary browsing tasks, participants in this group also needed to listen to the auditory output that might be presented at any time during the experiment. Figure 1 shows a screen shot of the “Visual + Auditory (Helpful Info)” mode.

Dependent variables included the number of correctly answered questions and satisfaction. User performance was measured by the number of correctly answered questions related to the text-based web site. Satisfaction was measured by a questionnaire based on original scale items developed by Davis (1989).

Twenty-two (22) participants have been recruited from a university in the U.S. Midwest region, which hosts a variety of different age groups, ethnicities, computer experience levels, and knowledge backgrounds. Participants were randomly assigned to the two groups: “Regular Visual Display” and “Visual + Auditory (helpful Info)”.

Each participant was asked to sign a consent form before participating in the experiment. Upon completion of a pre-experiment questionnaire, each participant received instructions for using the experiment browser and performing the requisite tasks. These instructions were presented via computer and in hard copy format. The participant then started a training session in which he or she could browse a sample web site with visual (and auditory) information similar to the version of the experiment the user will encounter. Upon successful completion of the training test, the participant began the experiment tasks. Considering that fatigue factor might affect participants' understanding capability and listening comprehension, the total duration of the experiment was limited to 30 minutes. All participants were asked to correctly answer as many questions as they could in the given time or less. Subjects were informed of the time limit prior to starting the experiment in order to focus their attention on the tasks at hand, and the program was programmed to automatically terminate at the 30 minute mark. Following the completion of the experiment tasks or the expiration of the experiment time allowed, each participant was 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 the “Visual + Auditory (Help Info)” group were then debriefed to provide additional information about their experience in processing and using the voice and text information. The debriefing sessions were recorded for data analysis.



(For “Visual + Auditory (Helpful Info)” mode only) Helpful hints.

Figure 1. Screen Shot of “Visual + Auditory (Helpful Info)” Presentation Mode

5. Results and Discussion

The intention of the hypothesis was to test the effectiveness and feasibility of a dual modal information presentation. It postulates that helpful information presented through auditory channel during a web browsing process can be perceived by users and will improve users' performance on browsing tasks. The dependent variables were the number of correctly answered questions and satisfaction. The descriptive statistics of each dependent variable for both experiment groups are presented in Table 1. The general satisfaction score was the sum of scores given to individual questions regarding the experiment Web site. Satisfaction score regarding the auditory presentation was the sum of scores given to the three questions about voice cues. In addition, the accuracy of task performance was also analyzed. It was defined as the division of the number of correctly answered questions over the total number of questions that the participant answered.

Table 1. Comparison between “Regular Visual Display” and “Visual + Auditory (Helpful Info)” Presentation

Variables	Regular Visual Display (n = 11)		Visual + Auditory (Helpful Info) (n = 11)	
	Mean	Std.	Mean	Std.
Number of correctly answered questions	20.7	10.43	26.9	7.20
Accuracy	0.742	0.1283	0.815	0.0755
General satisfaction	51.2	15.42	54.5	10.89
Satisfaction regarding auditory presentation	N/A	N/A	14.4	1.11

The preliminary results in Table 1 show that the “Visual + Auditory (Helpful Info)” group had higher mean values than the “Regular Visual Display” group in all three variables: number of correctly answered questions, accuracy, and general satisfaction. However, t-tests between the two groups suggests that the differences were not significant at the $\alpha = 0.05$ level (number of correctly answered questions: *t(20) = -1.62, p = 0.1215*; accuracy: *t(20) = -1.62, p = 0.1210*; general satisfaction: *t(20) = -0.56, p = 0.5822*). Although the hypothesis was not supported statistically, the "Visual + Auditory (Helpful Info)" presentation did show some improvement in user's performance in the browsing tasks.

Interestingly, another t-test indicates that the satisfaction score regarding auditory presentation in the "Visual + Auditory (Helpful Info)" group was almost significantly greater than score 12 (*t(10) = 2.14, p = 0.0583*). Based on the likert scale, score 4 (or 12 for the total score of 3 questions) means neutral. This result suggests that participants were somewhat satisfied with the auditory presentation and thus favors the hypothesis.

The short debriefing interviews conducted for the "Visual + Auditory (Helpful Info)" group at the end of the experiment reflect the following facts:

- Participants could recognize and remember the key words and phrases contained in the voice messages.
- The voice information did help participants perform the tasks.

Therefore, the hypothesis was partially supported by the preliminary results of this experiment. The results agree with prior research findings. Based on the multiple-resource human attention model proposed by Wickens (1980 & 1984), two tasks can be performed together more efficiently to the extent that they require separate pools of resources, such as different modalities. While users are visually browsing information, they may be able to receive brief information from the auditory channel.

6. Conclusions

In this study, a dual modal information presentation was proposed and tested through a controlled experiment. The findings from this study suggest that users might be able to perceive and use information presented through auditory channel while they are visually browsing textual information. Hence, it is possible to use multimodalities (visual plus auditory modes) to present more information than a single modality could. These findings could be applied in the multimodal interface design of mobile applications. However, this study is just the first step towards this direction. More participants will be needed to prove the significance of improvements. Future study must also address what information should be presented in the visual mode and in the auditory mode.

7. References

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.

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.

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.

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

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.

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.

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

Gorenflo, C. & Gorenflo, D. (1994). Effects of synthetic voice output on attitudes toward the augmented communicator. *Journal of Speech & Hearing Research, 37*(1), 64-68.

Krell, M. & Cubranic, D. (1996). V-Lynx: bringing the World Wide Web to sight impaired users. In *Proceedings of ASSETS'96,* 23-26, New York, NY: ACM.

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

Oviatt, S. & Olsen, E. (1994). Integration themes in multimodal human-computer interaction. In *Proceeding of the International Conference on Spoken Language Processing, 2,* 551-554, Acoustical Society of Japan.

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

Schlosser, R., Belfiore, P., & Nigam, R. (1995). The effects of speech output technology in the learning of graphic symbols. *Journal of Applied Behavior Analysis, 28*, 537-549.

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

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

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

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