Evaluating Usability Improvements by Combining Visual and Audio Modalities in the Interface

Carlos Duarte, Luís Carriço, and Nuno Guimarães

LaSIGE – Faculty of Sciences of the University of Lisbon
Edifício C6, Piso 3, Campo Grande
1749-016 Lisboa, Portugal
{cad, lmc, nmg}@di.fc.ul.pt

Abstract. This paper reports the findings of an evaluation of an adaptive multimodal application for reading of rich digital talking books. Results are in accordance with previous studies, indicating no user perceived difference between applications with and without adaptivity. The NASA Task Load Index was also used and showed that users of the adaptive application reported less workload. Results also include a comparison between tasks executed with electronic support and tasks executed with print support, and also what specific features in the interface benefited the most from the use of visual and audio modalities.

Keywords: Evaluation, Adaptive Interfaces, Multimodal Interfaces, Electronic and Print Reading, Digital Talking Books.

1 Introduction

In today's ever changing interaction scenarios, alternative modalities to the predominant visual channel, will have to be explored in order to cope with the new challenges. For interfaces to be flexible enough to adapt to these requirements, other modalities will have to be employed, either by themselves or in combination with visual interaction. Balancing the load between modalities, or even switching to different modalities, in order to better adapt the interaction characteristics to the information being presented, can decrease the cognitive effort required [1]. This interaction adaptation may improve the usability for the average user, and will certainly improve the accessibility for people with disabilities.

One of the most promising modalities for complementing or replacing visual interaction is audio. For the average user, sounds and speech are natural input and output modalities. For visually impaired users they are, perhaps, the most important input and output modalities. Research has proved that audio, used either as input or output, can provide solutions to specific interaction problems [2], but also that audio is more suited to specific tasks, while other tasks are better accomplished using other modalities [3]. From these results we can expect that audio related benefits can be optimized when combined with other modalities. Speech recognizers and voice synthesizers, whose performance has increased over the years, are beginning to be

deployed in general public applications, meaning more and more users have had contact with some kind of speech technology. However, most of these applications, like call centers, rely solely on audio. The combined used of two modalities remains outside of the general public reach.

In this paper we explore usability issues in an application using video and audio as input and output modalities. The following section briefly introduces the application used in the evaluation sessions. The next section describes the experimental setting and procedures. This is followed by the presentation of the evaluation results. Section 5 discusses the results, and the final section concludes the paper and presents future work.

2 Rich Book Player – An Adaptive Multimodal Digital Talking Book Player

The application used in the usability evaluation was the Rich Book Player, an adaptive multimodal Digital Talking Book player [4]. This player can present book content visually and audibly, in an independent or synchronized fashion. The audio presentation can be based on previously recorded narrations or on synthesized speech. The player also supports user annotations, and the presentation of accompanying media, like other sounds and images. In addition to keyboard and mouse inputs, speech recognition is also supported. Due to the adaptive nature of the player, the use of each modality can be enabled or disabled during the reading experience.

Figure 1 shows the visual interface of Rich Book Player. All the main presentation components are visible in the figure: the book's main content, the table of contents, the figures panel and the annotations panel. Their arrangement (size and position) can be changed by the reader, or as a result of the player's adaptation. The other visual component, not present in figure 1, is the search panel. Highlights are used in the main content to indicate the presence of annotated text and of text referencing images. The table of contents, figures and the annotations panels can be shown or hidden. This decision can be taken by the user and by the system, with the system behavior adapting to the user behavior through its adaptation mechanisms. Whenever there is a figure or an annotation to present and the corresponding panel is hidden, the system may choose to present it immediately or may choose to warn the user to its presence. The warnings are done in both visual and audio modalities.

All the visual interaction components have a corresponding audio interaction element, with one exception. Since the speech recognizer currently used in the player[1] does not support free speech recognition, annotations have to be entered by means of a keyboard. All the other commands can be given using either the visual elements or vocal commands.

[1] This applies to the Portuguese version of the player, which was the one used in the usability study.

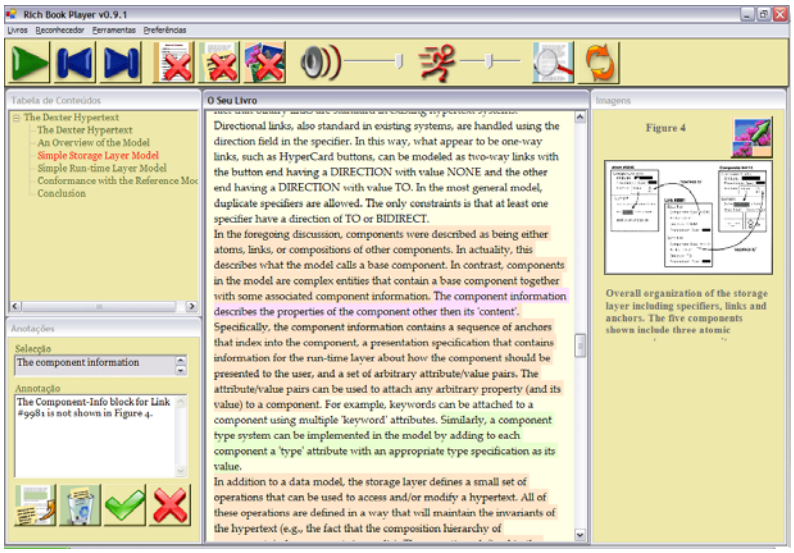

Fig. 1. The Rich Book Player's interface. The center window presents the book's main content. On the top left is the table of contents. On the bottom left is the annotations panel. On the right is the figures panel.

3 Experimental Setting

The usability evaluation was carried out in the context of an article reviewing assignment for a Hypermedia Systems course. The students had several such assignments over the semester, which consisted of preparing a summary and an oral presentation of a given article. The summary and the oral presentation were group tasks, typically done over a two weeks period. With the students' agreement, it was decided that one of those assignments was to be done with support from the Rich Book Player, over a one day period. The assignment consisted in reading the article "The Dexter Hypermedia" individually during the morning period, and preparing a group summary and answering a short test during the afternoon. Over a period of four days, thirty-three students participated in the evaluation: six in the first day, and nine in each of the other days.

Giver the number of simultaneous participants, and the length of each session, the experiment was not conducted in our regular usability evaluation laboratory, but a special setting was prepared in another room. The room was set up with nine test stations. Each station consisted of a laptop computer with a larger screen, mouse, headphones, microphone and webcam attached to it. The Rich Book Player application was available in all stations. The application was endowed with logging capabilities, thus recording all interaction with the participants. The stations also did screen recording, voice inputs recording, and webcam recordings, thus allowing for a

full backup of the experiment. In addition to the stations, two digital video cameras recorded other aspects of the interaction.

The experiment was divided in two periods. The morning period started with a 30 minutes period for application familiarization, which was followed by 120 minutes for article reading, and ended with a usability questionnaire. The afternoon period was composed by a 75 minutes session for summary preparation, 30 minutes for answering a short test without access to the article, 30 minutes for the same test with access to the article, and finally, another questionnaire. For the summary preparation task, the annotations of all the group's members were merged, and the group worked on only one station.

In order to be able to evaluate the effects of using a multimodal application on the task of reading an article, the students were divided in two major groups. The control group read the article printed in paper, and the test group read the article using the Rich Book Player. In order to investigate the effects of adaptation, the test group was further divided in two groups: a group with some of the adaptation features turned off, and other group with all the adaptation features on. In total the control group counted nine elements, and the other two groups, twelve elements each.

To reduce the effect of extraneous variables, the following controls were applied:

- The tasks were the same for each participant.
- The tasks had the same time constrains for all participants. The questionnaires were answered immediately after task completion.
- All test stations were equipped with laptop PCs of the same model (Sony VAIO TX3) and external monitors with the same dimensions. All stations were configured to use the same screen resolution, operating system version, applications and desktop configuration.

4 Evaluation Results

The experiment results consist of qualitative data, gathered from the different questionnaires answered by the participants, and quantitative data, gathered from the logs and screen and video capture. In this paper we present and analyze the results from the qualitative data.

Three sets of questionnaires were answered during the experiment by the participants from both test groups, and one set only by participants from the control group.

4.1 NASA Task Load Index

The first questionnaire administered to the participants was the NASA Task Load Index (NASA-TLX) [5]. All the participants answered this questionnaire since it focused on the task, not the application. The questionnaire was presented to the participants immediately after the completion of the article reading task.

The NASA TLX is a subjective workload assessment measure. NASA-TLX is a multi-dimensional rating procedure that derives an overall workload score based on a

weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration.

The NASA TLX was used in this experiment with the main goal of finding a difference between the scores of participants in the adaptive and non-adaptive groups, and between these groups and the control group. Previous findings [6,7] show users do not perceive advantages in using adaptive interfaces over non-adaptive interfaces. Using a subjective workload assessment measure might reveal a difference not directly perceived by the participants, leading to the following hypotheses:

H1 Performing the article reading task with the adaptive application, the non-adaptive application, or with a paper article, will result in different perceived workload measures.

Measures were collected for all participants (12 in the adaptive group, 12 in the non-adaptive group and 9 in the control group). A one-way ANOVA test was performed, and revealed that the perceived workload by users of the adaptive application (M = 53.30, SD = 14.27), users of the non-adaptive application (M = 57.11, SD = 13.45), and users with only a paper article (M = 57.56, SD = 14.79) did not differ significantly $F(2, 30) = 0.31$, $p > 0.05$. The statistical analysis does not support hypotheses **H1**, meaning that the perceived workloads do not differ significantly based on the support used for reading the paper.

4.2 Usability Questionnaire

Following the NASA TLX, participants in the adaptive and non-adaptive group were asked to answer to a second questionnaire. This 26 questions questionnaire focused on feature usefulness and application usability, and was organized in the following groups: Navigation, Annotations, Images, Search, Adaptation (only for the adaptive application group), Presentation, Interaction and General Opinion. All the questions were answered in a 10 point scale.

The General Opinion was measured on three questions, evaluation the participants' opinion and reaction to the application (figure 2).

The correlation between the answers to the three questions was calculated, and all three showed to be significantly correlated ($p < 0.001$). Taking this significant correlation into account, it was possible to reach a single measure of opinion by adding the answers to the three questions for each participant. In accordance to what has been presented before, no significant difference was expected to be found between the two groups, which lead to formulating the following hypotheses:

H2 The general opinion of users of the adaptive application is similar to the general opinion of users of the non-adaptive application.

To evaluate this hypotheses a *t-test* was performed on the data, showing that the opinion of people in the adaptive group (M = 18.92, SD = 6.05) was not significantly different from the opinion of non-adaptive group (M = 17.33, SD = 5.02), $t(22) = 0.70$, $p > 0.05$.

For each of the other question groups in the questionnaire, *t-tests* were applied to the usability related questions, in order to understand how the use of multimodal output (visual and audio combined) contributed to the overall usability of the application. In the following paragraphs all reported *t-tests* take into consideration the necessary Bonferroni adjustment.

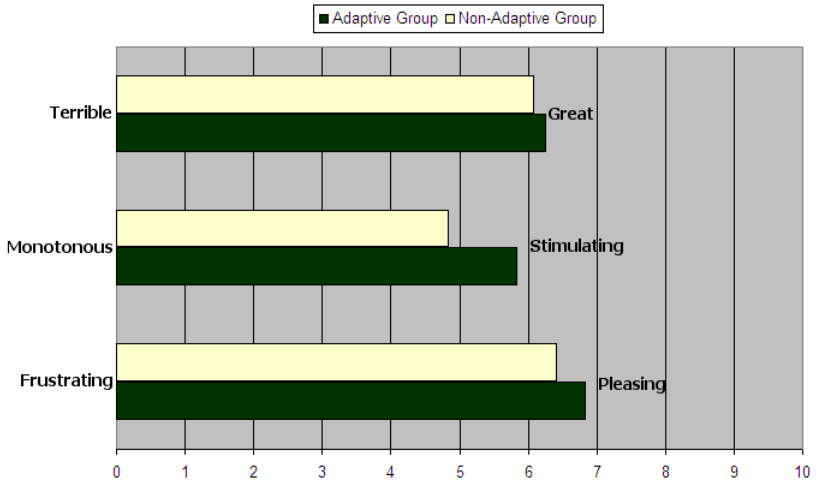

Fig. 2. Average of the answers per participant group to the three criteria on the General Opinion group of the usability questionnaire.

Regarding the navigation in the Rich Book Player, several features were offered, including navigation using the table of contents, going forward or backwards a word, sentence, paragraph or chapter, and by direct selection in the main window content. The results indicate it the available mechanisms were considered usable, $t(23) = 10.79$, $p < 0.001$.

Annotation creation is one of the most difficult mechanisms to implement. Previous evaluations showed it [8], and prompted an alteration of the steps necessary to create an annotation. This procedure was redesigned, making more explicit the need to first select the part of text being annotated, and only after that step inputting the annotation. Better support for text selection was developed, including an initial suggesting of the current sentence, and simple commands to expand this selection. However, both the sequence of commands to create an annotation, $t(23) = 1.79$, $p > 0.05$, and the commands for helping with the text selection, $t(23) = 2.00$, $p > 0.05$, did not reach statistical significance, meaning test participants did not consider them particularly usable.

Search results appear highlighted in the text. To improve context acquisition, the whole sentence where the search term exists is also highlighted with a different color (one lighter than the one used for highlighting the searched terms). This feature was considered to improve the usability, $t(21) = 4.21$, $p < 0.05$.

The application also tried to minimize the movement of the main text windows whenever another window appeared or disappeared from the screen, by controlling the appearance point, the width of the windows, and the position of remaining windows whenever a window was hidden. This feature was considered useful by the test participants, $t(23) = 4.20$, $p < 0.05$.

On an overall interaction rating, the Rich Book Player was considered usable by the participants, $t(23) = 7.05$, $p < 0.001$.

The awareness raising mechanisms made special use of the two modalities available, displaying text which had been annotated, or had an image associated with it, in different background colors, and also using verbal cues to signal the presence of such text. Current chapter was also highlighted in the table of contents, and after arriving at a new chapter, verbal cues indicated its number and name (whenever applicable). A series of questions concerned this features, and tried to evaluate if they helped the users become aware of their place in the book, and what content existed around their current reading point. All the answers showed these features to be usable and effective awareness raising mechanisms, $p < 0.05$.

4.3 Comparing Electronic and Paper Reading

The final questionnaire, presented after the group summary writing task, asked the participants from adaptive and non-adaptive groups to compare their experience of reading an article with the Rich Book Player application to that of reading printed articles. A questionnaire with eight questions comparing different aspects of the reading experience was prepared. Answers were given on a 5 point Likert Scale. Once again, all the *t-test* results presented in the following paragraphs have taken into account the necessary Bonferroni adjustment.

The first question compared navigation in the electronic format to the printed format. The average of the answers was 3.79, and a *t-test* revealed that participants felt navigation in the electronic format was significantly easier than in the printed format, $t(23) = 4.98$, $p\ 0 < 0.05$.

The next question compared searching in both formats. Answers' average was 3.96, and a *t-test* confirmed that participants felt that finding text in the electronic format is significantly easier than in the paper format, $t(23) = 4.7$, $p < 0.05$.

The two following questions deal with annotation creation and annotation reading. Neither of these showed statistically significant results. Answers for easiness of annotation creation were 3.00 in average, while for annotation reading 3.46 on average.

The next question dealt with how easy it was to acquire the context of an image in both formats. Once again the answer is not statistically significant, even though the average answer, 3.21, is above the scale's mid-point.

Questions six and seven dealt with which format did the users felt it was quicker to read, and easier to understand the article's contents. The average for the first one was 3.04, and for the second one 3.13, with both failing to reach statistical significance.

The last question asked which is the less tiring format for reading the article. Average answer was 3.08, not reaching statistical significance.

5 Discussion

The analysis of the experiments results conducted so far allows drawing some conclusions regarding the usage of a digital book player endowed with multimedia and adaptive features: the comparison of an application with adaptive features turned on and off, the comparison of performing a task with electronic or printed support, and the improvements in usability gained from combining two modalities (in this case, video and audio).

5.1 Adaptive versus Non-Adaptive Applications

When evaluating adaptive systems, additional problems have to be dealt with, in comparison to the evaluation of non-adaptive systems:

- The definition of a control group is difficult for those systems that cannot switch off the adaptivity to make a non-adaptive version, because it is an inherent feature of the system [9]
- Criteria for definition of adaptivity success are not well defined. On the one hand, objective standard criteria regularly failed to find a difference between adaptive and non-adaptive versions of a system. On the other hand, subjective criteria, standard in HCI research have been rarely applied to evaluation of adaptive systems [10].
- The effects of adaptivity in most systems are expected to be rather subtle in comparison to what may be expected from individual differences, and thus require precise measurements, potentially taking into account behavior and cognitive aspects of the users [11].

This study tried to deal with some of these aspects. By having some of the participants work with a print version of the article, it was possible to define a control group applicable to both adaptive and non-adaptive versions of the application. Furthermore, it was possible to turn off some of the application's adaptive features without rendering it unusable, enabling a comparison between two versions of the application. The study also tried to establish a comparison between adaptive and non-adaptive versions of the same application using different subjective measures.

The results, however, are in accordance to previous results in the literature, indicating no significant perceived differences between the adaptive and non-adaptive versions of the application, even tough the opinion of the participants who worked with the adaptive version of the application was, on average, higher than that of the participants who worked with the non-adaptive version.

The same can be said about the perceived workload measured by the NASA TLX, where, once again, no statistical significance was found in the results. In this case, the comparison extended to the participants working with the print version, who achieved scores very similar to those of the participants working with the non-adaptive version of the application. The participants of the adaptive application group achieved lower scores on the NASA TLX, indicating a lower perceived workload, even tough not enough to be statistically significant, but justifying further studies to investigate if this indicator can identify a difference between adaptive and non-adaptive applications.

5.2 Reading in Electronic and Print Supports

Another aspect evaluated in this study was the participants' opinion regarding the task of reading an article using an electronic medium offering multimodal output, compared to reading printed works.

A somewhat surprisingly result was the average answer to all the questions being above or equal to the 5-point Likert scale's medium point, meaning that no task was more difficult to perform in the electronic medium than in the printed medium. This was the expected result for some tasks, like searching, but not for other tasks like annotation creation.

However, only two tasks were significantly easier to perform with the Rich Book Player than with printed articles: navigating and searching. While this was an expected result for searching tasks, given the digital supports advantage, it is worth mentioning that navigation tasks also achieved the same level in the participant's opinion. This is probably explained by the vast possibilities offered for navigation inside the application, allowing users to navigate to any point with ease.

5.3 Improvements from Multimodality

Multimodal output is used throughout the application: content is presented visually and aurally, awareness raising mechanisms combine both modalities, and reading position is presented in both modalities also. Usability questionnaires assessed how the use of multimodality impacted the participants' opinion of the application.

The results show that combining visual and audio led to improvements not felt in other areas of the interaction, where the modalities were not used in combination. This was particularly felt in the participants' opinion of the usability of the awareness raising mechanisms.

6 Conclusions and Future Work

This paper presented the results of an evaluation of an adaptive multimodal Rich Digital Talking Book Player. This player combines visual and audio modalities, both for input and output, and is also endowed with adaptive capabilities, leading to the interface's behavior adaptation in response to changes in the user's behavior. The evaluation experiment counted with the involvement of 33 participants, arranged in three groups: an adaptive application group, a non-adaptive application group, and a control group which worked with printed texts.

Evaluation results confirmed no perceived differences between adaptive and non-adaptive applications. However, when considering the NASA Task Load Index, the workload felt was smaller for the adaptive application group. This result did not reach statistical significance, but nevertheless prompts the need for further experiments. When comparing tasks performed with the Rich Book Player, and tasks performed with printed texts, the participants' general feeling was that it was easier to perform tasks with electronic support. While for some tasks (e.g. searching) this was expected,

for other it was somewhat surprising. The use of multimodality has also proven beneficial from the usability viewpoint, particularly for implementing awareness raising mechanisms.

To gather further results that may shed some light on the effects felt with long term usage of an adaptive application, another experiment is currently underway, where the participants have the Rich Book Player at their disposal in their home environment for a period of two months.

References

1. Kalyuga, S., Chandler, P., Sweller, J.: Managing split-attention and redundancy in multimedia instruction. Applied Cognitive Psychology, Vol. 13, N. 4 (1999) 351–371
2. Sawhney, N., Schmandt, C.: Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. ACM Transactions on Computer-Human Interaction, Vol. 7, N. 3. ACM Press, New York (2000) 353–383
3. Oviatt, S., Coulston, R., Lunsford, R.: When Do We Interact Multimodally?: Cognitive Load and Multimodal Communication Patterns. Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA. ACM Press, New York (2004) 129–136
4. Duarte, C., Carriço, L.: A conceptual framework for developing adaptive multimodal applications. Proceedings of the 11th International Conference on Intelligent User Interfaces, Sydney, Australia. ACM Press, New York (2006) 132–139
5. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: P.A. Hancock, N. Meshkati (eds.): Human mental workload. North-Holland, Amsterdam (1988) 139–183
6. Höök, K.: Evaluating the utility and usability of an adaptive hypermedia system. In: Moore, J., Edmonds, E., Puerta, A. (eds): Proceedings of the 2nd International Conference on Intelligent User Interfaces, Orlando, Florida, United States. ACM Press, New York (1997) 179-186
7. Weibelzahl, S.: Evaluation of Adaptive Systems. PhD Dissertation. University of Trier, Germany (2003)
8. Duarte, C., Chambel, T., Simões, H., Carriço, L., Santos, E., Francisco, G., Neves, S., Rua, A.C., Robalo, J., Fernandes, T.: Avaliação de Interfaces Multimodais para Livros Falados Digitais com foco Não Visual. Proceedings of the 2nd Conferência Nacional em Interacção Pessoa-Máquina, Braga, Portugal (2006)
9. Höök, K.: Steps to take before intelligent user interfaces become real. Interacting with computers, Vol. 12, N. 4. Elsevier (2000) 409-426
10. Weibelzahl, S., Lippitsch, S., Weber, G.: Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. Künstliche Intelligenz, Vol. 16, N. 3 (2002) 17-20
11. Karagiannidis, C. Sampson, D. G.: Layered evaluation of adaptive applications and services. In: Brusilovsky, P., Stock, O., Strapparava, C. (eds.): Proceedings of Adaptive Hypermedia and Adaptive Web-Based Systems, International Conference, Trento, Italy. Springer, Berlin (2000) 343-346