Usability Evaluation of Digital Talking Books

Carlos Duarte, Luís Carriço
LaSIGE, Dep. de Informática da Faculdade de Ciências da Universidade de Lisboa
Campo Grande, Edifício C6, 1749-016 Lisboa
{cad,lmc}@di.fc.ul.pt

Abstract

In this paper we present the results of a set of usability evaluation studies on Digital Talking Books. Digital Talking Books aim to provide better access to literary content for the blind and visually impaired. Departing from their digital base, we developed enriched books, targeting broader audiences and usage situations. The books can be enriched with different media content, like images and sounds, supporting multimodal interaction and multimedia presentation. Enlarging the number of interaction possibilities, while making the books more attractive, also increases the possibility of usability problems. This is why it is so important to perform usability evaluation of the generated interfaces. A special focus of the evaluation was given to the synchronization related issues. Synchronization flaws detected by the users were the cause of discomfort and loss of concentration. Alternative synchronization visual marking was employed to circumvent these problems. Tests also provided insight on what modalities are favoured by the users for specific interaction tasks, and confirmed the importance of the availability of multimodal interaction.

Keywords

Usability, Evaluation, Multimodal interfaces, Digital talking books.

1. INTRODUCTION

Audiotapes have been the most widely used means of access to literary content by the blind and print-disabled community. Even with all the merits that should be attributed to them, several lacking functionalities and usability problems can be easily identified. For example, searching for an expression might involve listening to the whole audiotape. Although simply stopping the tape and resuming it later in the same point creates an implicit and temporary bookmark, the creation of additional bookmarks is a more demanding task. And when considering annotating the book or following cross-references, we are facing cumbersome or impossible tasks.

The introduction of the Digital Talking Book (DTB) is an attempt to bring these and other functionalities to the blind and print-disabled communities, by joining the written and spoken words in a digital format. In a DTB the reproduction of a digital recording of the narration may accompany the reading of the text. The presence of the source text in a digital format makes navigating the book, searching for words and other possibilities available. Worthy of mention is also the improvement in the reproduction quality insured by the digital recording of the narration.

The introduction of different digital media, indexed to the source text, combined with a multimedia playback environment, affords the expansion of the traditional reading experience, reshaping it into an "immersion" in a multimodal environment. The possibility of creative combination of presentation elements, taking advantage of available media resources, offers the support to new ways of telling stories and improving learning [Carriço03]. The evolution is based on the introduction of new multimedia elements, in a coherent way, during "reading". Possible enrichments are the introduction of background music, environmental sounds related to the "scene of the action", images or videos to complement information presented in the original source, and many more. Here, as with traditional DTBs, the use of an automatic and flexible production platform, allowing the creation of enriched books, it of the utmost importance.

Besides the visually impaired community, who is the main target audience for DTBs, several other population segments can benefit from a platform with these characteristics. The multimodal interaction capabilities of the books broaden the reading opportunities to situations where the reader is engaged in other visual activities, such as driving or surveillance, allowing the user to issue voice commands and supporting features that could not be made available by a simple digital storage medium (like a CD or DVD). Balancing DTB modes and media can be explored to overcome the cognitive limitations of human perception and attention [Gazzaniga98].

The interaction possibilities for such a book are greater than the ones for standard DTB players [Dolphin03] [IRTI03], but there is also a greater number of possible usability problems, caused by the added capacities and the use of different media [Duarte03]. It is therefore a necessity to perform usability evaluation of the books' interfaces, to guide the development process, and to al-

low for a better fit of the interfaces to the several user groups and usage conditions in which DTBs are expected to perform. This is precisely the focus of this paper.

In the next section an overview of related work is presented, covering DTB standards and navigation features, and non-visual interfaces. Next, the DTB building process is briefly described. This is followed by a description of the characteristics of some of the interfaces created so far and the usability issues related to the presentation, synchronization, navigation and enrichment of the interfaces. We then present a report of the evaluation studies conducted, their results and their impact on the development of the interfaces. We finish by drawing some conclusions and presenting ongoing and future work.

2. RELATED WORK

2.1 Digital Talking Books

DTBs are intended to provide an easier access to books for the blind and print-disabled community. Members of those communities cooperated with several organizations that developed DTB related standards. In Europe, the Daisy Consortium, with collaboration from the European Blind Union developed one of those standards. In the USA the National Information Standards Organization (NISO) in collaboration with The National Library Service for the Blind and Physically Handicapped conducted a similar work. As a result several DTB specifications have been proposed and evolved over the last years. Recently a joint effort of these organizations resulted in the most important DTB specification, the ANSI/NISO z39.86 [Ansi/Niso02].

DTBs can be classified according to the presentation and interaction possibilities made available, which also reflect their inherent complexity [Daisy02]: full audio with title only; full audio and navigation control; full audio, navigation control and partial text; full audio and full text; full text and partial audio; full text and no audio. Our DTB generation framework supports the creation of the more complex class, full audio and full text, but nevertheless, all the DTB classes can be generated within the framework.

According to the NISO Document Navigation Features List [Niso99], a DTB should provide basic navigation capabilities (advancing one character, word, line, sentence, paragraph or page at a time, and navigation to specific segments of the DTB), fast forward and reverse, reading at variable speeds, navigation through table of contents or a navigation control file (allowing the user to obtain an overview of the material in the book), reading notes, cross-reference access, book marking, searching and others.

However, and wisely, no specific implementation solutions are present in the standard. The solutions must consider aspects related to the proposed specification, but also the non-visual nature of the targeted environment.

2.2 Speech Interfaces

The work on non-visual interfaces can provide us with clues on how to tackle some of the problems faced.

Voice browsers are devices that exhibit at least one of the following characteristics: (1) can render web pages in audio format; (2) can interpret speech for navigation. Voice browsers and DTB interfaces share some common problems:

- The audio format is a temporal medium. A visually presented page can render simultaneously images, tables and text, in a spatial format, which is quickly processed by the perceptual human system. Spoken text, however, can present only one word at a time.
- Issuing voice commands, and audio processing, are activities that consume working and short-term memory, conflicting with planning and problem solving tasks. Visual information is processed by separate cognitive systems [Christian00].
- The unavoidable recognition errors.

However, the research in the multimodal systems field have made it clear that speech input is advantageous under certain circumstances [Oviatt00]. Studies [Van Buskirk95][Christian00] point out "the best tasks for speech input were tasks in which the user has to issue brief commands using a small vocabulary".

The interaction characteristics of a DTB are advantageous for the adoption of a speech interface: a relatively small number of commands can be used to implement the needed functionalities. However, some limitations may arise, if, for instance, to follow a table of contents entry, the user is asked to speak the chapter's title.

Research on the efficiency of speech as an input mode is not conclusive, although showing an increase in task completion time [Van Buskirk95][Christian00].

Some of the recommendations made for constructing voice browsers can be adopted for the design of DTBs:

- Links should be easily spoken text.
- Links should be short (a few words).
- Avoid links with similar sounds.
- Develop alternatives to numbered links, as these cause cognitive overload.

Methods for conveying document structure and assist in the navigation, in a non-visual environment, have also been researched. The use of 3D audio is proposed in [Goose99]. To convey document content, such as the presence of links and headings, the use of auditory icons [Gaver93][Blattner90], multiple speakers and sound effects [James97], amongst other techniques have been studied.

3. DIGITAL TALKING BOOK PRODUCTION

The section gives a brief overview of the DiTaBBu platform [Carriço04a][Carriço04b] used to build the DTBs.

The DTBs are built from digital copies of the source text and the audio narration. The framework separates the building process in two phases: content organization, and interface generation. The separation of the content from the interface facilitates the creation of different interfaces

for the same book. Besides the synchronization and navigation issues specific to the books, the framework also considers the playback platform and environment, the interaction devices available and the characteristics of the users, when generating the book's interface.

The building platform has the ability to generate books in several formats for which a set of templates is available. Templates for visual presentations, audio input, and audio output have already been developed for presenting the books in a PC web browser, and more recently on a SMIL player. For the web browser presentation templates exist that are capable of producing the output in two different languages: HTML+TIME [Microsoft02], which is Microsoft's implementation of SMIL [Bulterman01], the synchronization language proposed in the ANSI/NISO standard; and HTIMEL [Chambel01], another multimedia synchronization language. The templates enable the production of audio only presentations by combining the audio input and output templates, and multimodal presentations with any combination of visual and audio modalities.

Figure 1 shows the general layout of an interface (examples are built on the novel "O Senhor Ventura" by Miguel Torga) using audio and visual modalities for both input and output.

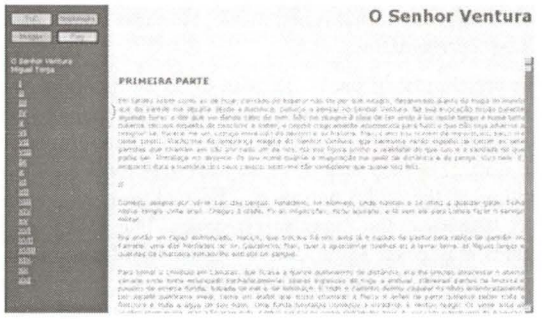

Figure 1 – Possible layout for a DTB interface, including Table of Contents.

4. DIGITAL TALKING BOOK FEATURES AND USABILITY CONCERNS

This section presents some of the features of the generated interfaces, and discusses issues related to the integration and synchronization of the different modalities.

Conveying to the user the difference between annotations, structural navigation and synchronization, without ambiguity, while following a sound interface metaphor, poses usability problems [Morley98]. Some of the problems are related to specific modalities, while others arise from the multimodal presentation of the book. However, the complementary use of different modalities can help us avoid some of the usability problems.

4.1 Synchronization

The audio narration of the book's content is synchronized with the visual display. Several synchronization units can be produced by the DiTaBBu platform. At the lower level - the word - the visual marking emphasis the word that is being narrated. The user has a very detailed notion of the synchronization between text and sound. However, the amount of information the system needs for synchronization is also the largest of all units, resulting in frequent losses of synchronization between the visual and audio modes, unless the book (or the HTML document to be more precise) is of small size. One way to improve the synchronization's accuracy with this unit is to partition the book into logical pages, which may or may not coincide with the physical ones. However, this has the disadvantage of causing the system to load pages more often, which might disrupt the reading experience.

Other synchronization units can be used. It is possible to have a synchronization unit for every syntactic construct present in the book: sentences, paragraphs, sections, etc. Another possible unit is called the silence, and it comprises all the words between two reader pauses. In comparison with the word unit, the silence has lesser synchronization detail. However, a greater level of accuracy for documents of comparable size compensates this. Besides, it is the most expected unit since the reading pace accompanies the hearing one.

Independently of the unit used, the synchronization can be presented to the user in one of two ways: by highlighting the words currently being read, or by accompanying the narration point with a side-margin marker. Figure 2 displays both ways of presenting the synchronization: highlighting in the top, and side-margin marker in the bottom.

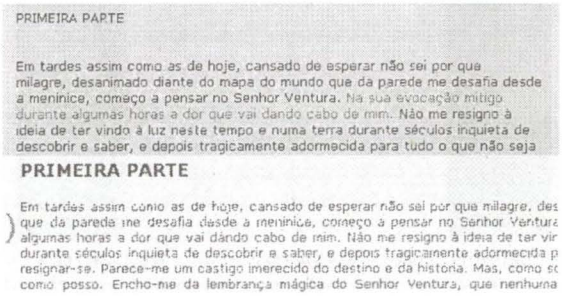

Figure 2 – Two ways of visually presenting synchronization for silence based synchronization. Top – highlighting; Bottom – side-margin marker.

Synchronization related issues are specific to multimodal environments, but some can be felt even when the output is only visual. This is the case of a presentation with text and images, where the images must still be coordinated with the text displayed. When the output is both audio and visual the synchronization problems faced are more challenging.

Higher-level synchronization, such as relating the current narration point to sections or chapters, can be transmitted visually by highlighting the corresponding entry in the table of contents. If there is no visual representation of the table of contents, this synchronization can be conveyed by speaking the chapter number, whenever a new chapter starts, or whenever a jump takes the narration to a different chapter.

In summary, two possible focus of usability problems exist related to the synchronization of the content. (1) If the synchronization unit is too low level, a synchronization loss can exist when the book is not "small" or not divided into smaller pages. In the case of using a too high-level synchronization unit it might not serve its purpose efficiently. (2) The visual presentation of the synchronization should guide the user to the text being narrated without distracting him from the reading.

4.2 Navigation

Displayed on the interface shown in figure 1 is the table of contents (to the left of the text area). Different levels of table of contents entries can be generated during the book production. The table can show entries ranging from the parts and chapters of the book (higher levels) to paragraphs (lower level), depending on the syntactic marking available. Usual table of contents entries are for parts, chapters and sections. The table of contents can be used to navigate the book, either through mouse clicks on its entries, or by issuing voice commands like "go to three", which would take the narration to the third chapter if the table of contents was displaying chapters.

Figure 3 shows another possible arrangement of the interface for the same book. This time the bookmarks and images tabs are displayed.

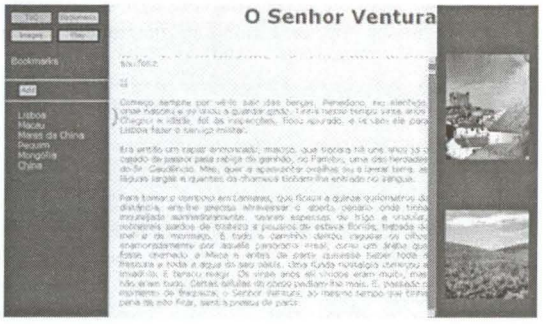

Figure 3 – Another possible layout for a DTB interface, including bookmarks and images.

To the left of the text the user created bookmarks and annotations can be seen. Similarly to the table of contents the bookmarks can be used to navigate the book through mouse clicks or voice commands. Clicking on an annotation, besides navigating the book by taking the narration to the annotation's creation point, also exhibits in a separate window the user created annotation's content. Even though most of the interacting might be done using mouse and keyboard or voice commands, annotation entering is currently only supported via keyboard. This is justified by the decrease in performance of the voice recognizer when not using a limited grammar. This limited grammar use is possible when recognizing commands for basic interaction with the book only, but impossible if trying to recognize annotations that could consist of any word.

Concerning the book navigation, besides the aforementioned navigation through the table of contents entries and created bookmarks, the user can also navigate the book by clicking anywhere on the text, or by issuing voice commands to move the presentation forward or backward by fixed amounts, paragraphs, sections and chapters.

Usability issues related to the choice of the destination point arise when considering the results of a search, or a navigation jump inside the book. The problem identified concerns the possible context loss suffered by the user when the beginning and end points of the jump are distant. To try to minimize this effect we propose the use of different jumps in the different modalities used. For example, after performing a search for a word where the first found occurrence is in the same section, the narration may resume on the searched word, but the visual display may jump to the start of the searched word's paragraph. If the first found occurrence is in a different section, the visual display may move to the beginning of the section, the narration may start on the searched word's paragraph, and the searched word may be highlighted.

4.3 Enrichment

To the right of the text (in figure 3) is presented the multimedia book enriching content. This may consist of pictures (as in figure 3), videos, links to available web pages, and others. It is also possible to enrich the book with background sounds and music, but these do not have visual representations.

The enrichment of the books with different media is also a possible focus of usability problems. The presentation of enhancing media can divert the user's attention from the story. The playback of video files may force an interruption of the audio narration. The inclusion of images may take away too many screen space. The background sounds may cover some auditory clues used for transmitting navigation or structural information. All these issues have to be considered when designing enhanced book presentations.

4.4 Presentation

A book can be presented in visual only format, audio only format, or using both formats. The absence of one of the formats can prevent the use of some of the media available. For example, in an audio only platform, images can only be presented in an alternative format, for instance, a previously recorded description of the image. In a visual only platform, background sounds cannot be conveyed to the user. When in presence of both output modalities we have to choose how to convey the information. Situations arise where it must be considered whether to use the different modalities in a redundant or complementary way [Martin98]. One example is the visual presentation of the book's content, accompanied by the audio narration. In this case the visual and audio modalities are used in a redundant way. Another example is a multimodal presentation in a PDA with audio narration, but no text, with the screen space used for visual display of the navigation structures (e.g. table of contents), thus allowing the user to navigate the book without needing to

use a voice recognizer module. In this case the visual and audio modalities are used in a complementary way.

Usability problems exist for both multimodal and non-multimodal interfaces. In the non-multimodal interface, they are not trivial when considering audio only presentations. In visual presentations, the concerns are primarily related to the screen disposition of the structures available to the user, and how to convey the relation between them. Traditionally, these structures are the book content, occupying the larger portion of the screen, and the table of contents, offering the possibility to quickly navigate to its entries. Other structures such as user made bookmarks and annotations, and the visual enriching content can also be present. To transmit relationships between these structures different possibilities can be employed. Highlighting the table of contents entry corresponding to the currently displayed chapter connects the table of contents to the text. Including visual markers in the text to signal the presence of a bookmark does the same for the bookmarks and annotations. Placing the images and videos aligned with the text paragraph they refer to, achieves the same result for the enriching content.

These same structures must be conveyed to the user in an audio only presentation, but the problem shifts from a location related one, to a temporal one. Our present approach, validated by interviews with print-disabled people, is to present the table of contents (and other tables if available) and the bookmarks before starting the playback of the text. Annotations, footnotes and side-margin notes are presented in their point of creation. Nevertheless, this information must be available and presented to the user any time it is asked for. Feedback on the chapter currently being read must also be made available as requested. The transmission of non-audio enriching content in this setting is more troublesome. Unless there are recorded verbal descriptions of the accompanying images and videos, that content cannot be transmitted to the listener.

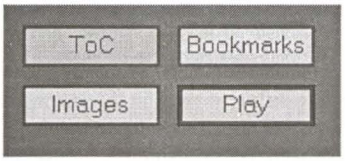

Figure 4 – The DTB interface control centre.

Also present in the visual interface is the control centre, shown in figure 4, and usually displayed in the top left corner of the screen. It allows the user to control what is displayed on the screen together with the book text. The screen content automatically rearranges itself when the user hides or shows one of the components. From here the user can also control the playback of the presentation, being able to pause and resume the narration. All these instructions can also be issued through voice commands.

5. USABILITY EVALUATION

This section presents the results from all the experiments we have conducted up to this point.

Three sets of tests have been conducted so far. The first two were designed to evaluate interfaces with all the interaction possibilities available. In these tests the goals were to identify general usability problems related to the navigation and presentation of the interface, and in particular evaluate how different ways of presenting the synchronization to the user influence the usability of the interface. The third set of tests was designed to evaluate a voice commands only version of the interface. Another goal of this test was to gather information to build an interaction grammar for a Portuguese speech recognizer. In this Wizard of Oz test the book was stripped of its English voice recognizer and one of the authors did the recognition work. In all the experiments audio and visual output was used. This means that the interfaces being tested were not designed for the visually impaired users, and as such, most of the results obtained cannot be applied directly to those interfaces. Nevertheless the results can provide a basis for the development of an audio only version of the interface. Table 1 summarizes the conditions of the tests.

Test #	Number Of Subjects	Sync. unit	Visual sync. presentation	Input modes	Output modes
1	8	Silence	Highlight	Mouse, Keyboard, Voice	Audio, Visual
2	8	Silence	Marker	Mouse, Keyboard, Voice	Audio, Visual
3	12	Silence	Marker	Voice	Audio, Visual

Table 1 – Summary of the evaluation tests conditions.

None of the subjects suffered any kind of visual or audio impairment. The interfaces used had no visual or audio enrichment, presenting only the text, the table of contents and previously created bookmarks. The only difference in the interfaces from test 1 to the other tests was the visual presentation of the synchronization between audio and text.

The concepts relating to the book's interface were explained to each of the subjects, and they were allowed as much time as needed to familiarize themselves with the interface. When ready they were handed a set of twelve tasks. The tasks were divided in two groups. The first group asked questions about the book content, providing hints to the location of the answer in the form of the chapter number or a close by annotation. A second group asked the user to find a specific point in the storyline, providing the same kind of hints, and to create an annotation in that point. These tasks required navigation either from the table of contents or from the annotations. In the first two experiments the subjects were free to choose how to issue commands. In the last experiment the subjects were constrained to using voice commands only.

Although the tasks are not representative of most of the navigation operations that would be done in a real setting (in the majority of the situations the user has no precise

hints about where the info he is looking for is located) they still allowed an evaluation of the navigational features ease of use.

After completion of the tasks each user answered a questionnaire. The questionnaire focussed on the perceived ease of use, utility and satisfaction levels of the users.

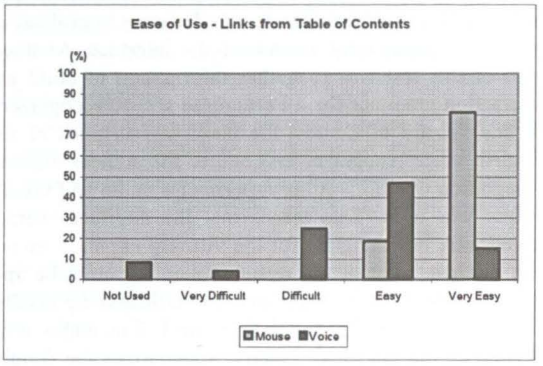

Figure 5 – Ease of Use of Input Modes when following links from the Table of Contents.

Figures 5 to 8 present the results for the ease of use of the input modes for each of the experiments. As can be seen, the mouse is consistently considered easier to use when compared to the voice commands, for navigation tasks (following links from the table of contents and bookmarks). This is easy to understand given that the accuracy of the voice recognizer when following links from the table of contents and bookmarks was 36% and 33.3% respectively. This forced the users to repeat the voice commands (or giving up on the voice commands), diminishing their ease of use. When taking into account the other operations it can be seen that the voice commands are considered easier to use. The voice commands for these operations are shorter, in accordance to the recommendations for voice browsers, so we can expect a better performance from the voice recognizer. For annotation creation operations the accuracy was 100%, and for playback control 97,2%, confirming the expectations. Note that when the performance of the voice recognizer is comparable to the mouse accuracy (which can be said to be 100%) the users find the voice commands easier.

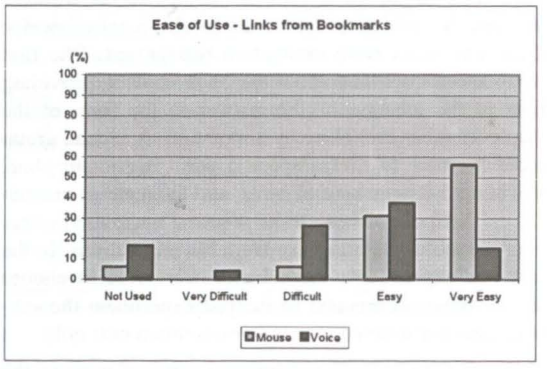

Figure 6 – Ease of Use of Input Modes when following links from the Bookamrks.

Figure 9 presents the utility of the available functionalities. Bigger relevance was attributed to the possibility of controlling the narration playback, followed by the navigation of the book through links from the table of contents. The other two functionalities were still considered very useful by the majority of the users.

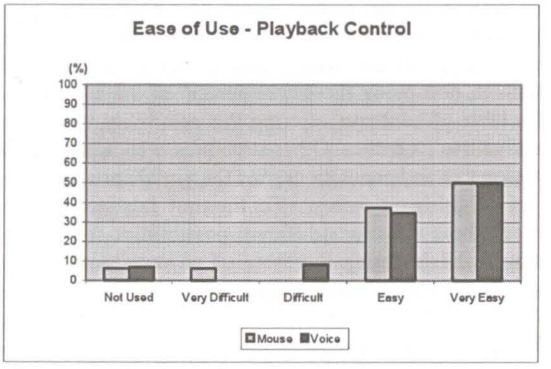

Figure 7 – Ease of Use of Input Modes when controlling the playback.

Users from the first and second experiments also evaluated the utility of the available multimodal interaction. It was considered indispensable by 25% of the users, very useful by 56.25% and of little utility by 18.75%.

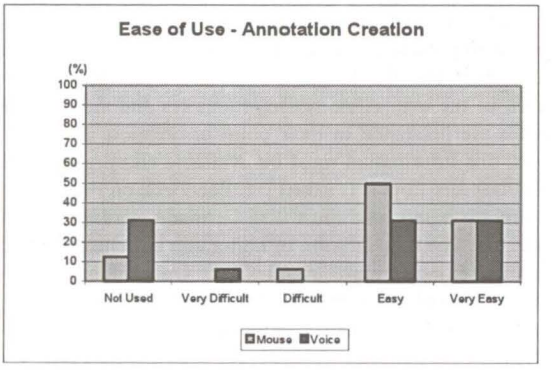

Figure 8 – Ease of Use of Input Modes when creating annotations.

Table 2 gives us an idea of the input modes preferred by the users for the different tasks in the second experiment. For navigation and annotation creation the majority of the commands were issued using the mouse, although not by a great margin. However, the users show a clearer tendency to use voice commands to control the playback of the book. This is consistent with the observed behaviour exhibited by several users during the evaluation, which tried to free their hands from the mouse, in order to have them available for writing down the answers to the questions presented as tasks.

Regarding the satisfaction levels, classified by the subjects of the first and second experiments on a scale of one to five (with five being the most satisfied), eleven users gave the mouse interaction the top grade, while only one user classified voice interaction with the highest mark. The worst grade for the mouse was a three given by one

user, while for the voice commands four users graded it with only a two.

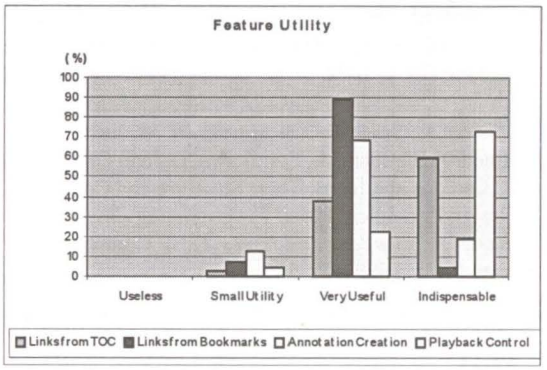

Figure 9 – Utility of available functionalities.

5.1 Discussion

Subject commentaries and test observations allowed us to identify some usability problems. Some of the problems (lack of feedback on the chapter currently being narrated, no numbering on the annotations) were easily solved. The solution to other problems is more demanding. The problems related to the performance of the voice recognition system, and the decrease in user confidence they originate, are being handled by replacing it. The new voice recognition system will support Portuguese, and is expected to employ different grammars for different interaction purposes. As the users that took part in the tests are native Portuguese speakers, with the exception of one, the requirement of having to use English commands imposed by the recognition system could have had an impact in the performance of the recognition engine. The new voice recognition system will support Portuguese, and employ a grammar designed for the basic interaction commands (navigation and playback control) constructed from user dialogues, and another for search operations built from the book contents itself. By separating the grammars it is expected an increase in the performance of the recognizer for each of the tasks.

Another important issue is related to how the synchronization between displayed and narrated text is transmitted to the reader. The conducted tests allowed us to infer the importance of a precise synchronization. The interface in the first test highlights words, while the interface for the second test uses a marker to accompany the text being narrated, but in both cases the synchronization unit is the silence. With the first interface, users reported discomfort when they perceived synchronization flaws. These caused users to loose concentration, making it difficult to accompany the text narration. Some users opted to switch off the narration and reading only the text. With the second interface, no such problems were reported. However, some of the second test users mentioned they would prefer a system showing them which word was being spoken.

The two main points to retain from these tests are: (1) the way to present the synchronization to the reader can "hide" synchronization flaws, eliminating the negative effects they have; (2) the compromise between accuracy and detail of synchronization varies from user to user, but further tests are needed to prove this point. A possible improvement to the interface will be to offer the user the possibility of choosing the synchronization unit to use and how to present the synchronization.

Task	Mouse	Voice
Links from TOC	53,7%	46,3%
Links from Bookmarks	55,6%	44,4%
Annotation Creation	55,6%	44,4%
Playback Control	37,9%	62,1%

Table 2 – Distribution of input modalities for the different tasks.

The wizard of Oz tests conducted with an interface accepting only voice commands, where one of the authors replaced the voice recognizer in order to allow the users to issue Portuguese voice commands, were motivated by two main goals: (1) gathering of the voice commands issued in an unconstrained situation of use, for the latter creation of the interaction grammar for the Portuguese voice recognizer; and (2) determining if the use of Portuguese instead of English contributed to an increased ease of use for Portuguese users.

Comparing the utility of the available functions of this interface with the ones tested earlier, more subjects considered indispensable the existence of links from the table of contents to the text (67% versus 56%) and the playback control (82% versus 69%). The same amount of subjects considered the links from annotations to the text very useful in both experiments (92%). The bigger importance given to the existence of the links from the table of contents to the text, and to the possibility of playback control, is justified by the absence of an easier way to skim through the book (which can be easily done with the mouse acting on a slider bar).

Regarding the ease of use of voice commands, the Portuguese version achieved better results than the English version, most notably for the link following tasks. Two possible explanations can be given: (1) besides being the native language of the users, Portuguese is also the language the book is written. The requirement of issuing English voice commands while reading text and listening to the narration in Portuguese can create complications for the users; (2) although random recognition errors were introduced during the wizard of Oz tests, the recognition performance was still better (in accordance to what we expect the recognizer to be able to achieve).

5.2 Preparing the Interface for Print-disabled Users

An interview with a print-disabled person, responsible for the audio books department of the Portuguese National Library, allowed us to identify some of the usability problems encountered by the visually impaired users of the National Library in audio books, and prepare a set of usability and accessibility tests to evaluate audio only

versions of our DTBs. Next we present some of the more relevant topics:

- All the information from the physical book should be available (including the cover). It should be presented in the same order as in the original book.
- References and annotations should be presented in the point where they appear in the text, and the listener should be made aware when a reference or annotation is being narrated. It is an improvement to be able to skip the listening of some of the references and annotations (listener choice). If possible, different narrators should read them.
- It should be possible to jump to the start of sentences, paragraphs and chapters.
- It should be possible to request to have any word spelled out. This is useful to learn how to write a word.
- An interpretation of tables and pictures should be narrated, instead of a narration of their visual content. This means that the narrator must not read the table line by line, but instead must provide the meaning of the table.

6. CONCLUSIONS

This paper presents the results of usability evaluation studies of Digital Talking Books. The books are created in a production framework, from digital copies of the source text and narration. The book presents the text and the narration synchronized. The reader can interact with the book through keyboard, mouse and voice commands, either independently or in a coordinated manner.

The usability tests revealed the importance of a precise synchronization between text and speech. Synchronization flaws caused user discomfort and loss of concentration. However, an interface that used a different synchronization presentation format, despite using the same synchronization unit, was able to eliminate the discomfort.

A poor performance of the voice recognition engine used was responsible for some of the usability problems encountered. For playback control, where the recognition engine performed well, the majority of the users adopted the use of the voice commands. And 81% of the users considered the possibility of multimodal interaction very useful or indispensable.

A new voice recognition system is being developed for the interface. This will support Portuguese language (the former forced users to issue commands in English), and by using a basic interaction grammar and a search grammar, is expected to improve the recognition performance.

Users' comments, test observations, and an evaluation of the characteristics of blind and visually impaired users, the main target audience for DTBs, made us aware of the necessity of making available personalized versions of the presentation and interaction. That prompts the development of adaptable books [Duarte04a]. We envision the creation of books tailored to a predetermined group of users, and books created for a general group, but with greater adaptation capabilities [Duarte04b].

7. REFERENCES

[Ansi/Niso02] ANSI/NISO, Specifications for the Digital Talking Book, 2002. <http://www.niso.org/standards/resources/Z39-86-2002.html>

[Blattner90] Blattner, M., Sumikawa, D., and Greenberg, R. Earcons and Icons: Their Structure and Common Design Principles. *Visual Programming Environments: Applications and Issues*. IEEE Computer Society Press, 1990, 582-606.

[Bulterman01] D. Bulterman. SMIL 2.0: Overview, Concepts, and Structure. *IEEE Multimedia*, 8(4), 2001, 82–88.

[Carriço03] Carriço, L., Guimarães, N., Duarte, C., Chambel, T., and Simões, H. Spoken Books: Multimodal Interaction and Information Repurposing. *Proceedings of HCII2003, International Conference on Human-Computer Interaction*, 2003, 680–684.

[Carriço04a] Carriço, L., Duarte, C., Lopes, R., Rodrigues, M., and Guimarães, N. Building Rich User Interfaces for Digital Talking Books. *Computer-Aided Design of User Interfaces IV*. Kluwer Academic Publishers, 2004.

[Carriço04b] Carriço, L., Duarte, C., Guimarães, N., Serralheiro, A. and Trancoso, I. Modular Production of Rich Digital Talking Books. *Proceedings of ICEIS2004*, 2004, Vol. 5, 158–163.

[Chambel01] Chambel, T., Correia, N., and Guimarães, N. Hypervideo on the Web: Models and Techniques for Video Integration. *International Journal of Computers & Applications*, 23(2), 2001, 90–98.

[Christian00] Christian, K., Kules, B., Shneiderman, B., and Youssef, A. A Comparison of Voice Controlled and Mouse Controlled Web Browsing. *Proceedings of ASSETS'00*, 2000, 72–79.

[Daisy02] Daisy Consortium. Daisy Structure Guidelines, 2000. <http://www.daisy.org/publications/guidelines/sg-daisy3/structguide.htm>

[Dolphin03] Dolphin Audio Publishing. EaseReader – the next generation DAISY audio eBook software player, 2003. <http://www.dolphinse.com/products/easereader.htm>

[Duarte03] Duarte, C., Chambel, T., Carriço, L., Guimarães, N., and Simões, H. A Multimodal Interface for Digital Talking Books. *Proceedings of WWW/INTERNET 2003*, 2003.

[Duarte04a] Duarte, C. and Carriço, L. Identifying Adaptation Dimensions in Digital Talking Books. *Proceedings of IUI'04*, 2004, 241-243.

[Duarte04b] Duarte, C., Carriço, L. and Simões, H. A Flexible Interface Architecture for Digital Talking

Books. *Proceedings of ICEIS04*. 2004, Vol. 5, 146–151.

[Gaver93] W. Gaver, Synthesizing Auditory Icons. *Proceedings of INTERCHI'93*, 1993, 228-235.

[Gazzaniga98] Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. *Cognitive Neuroscience - the Biology of the Mind*. W. W. Norton & Company, 1998.

[Goose] Goose, S., and Moller, C. A 3D Audio Only Interactive Web Browser: Using Spatialization to Convey Hypermedia Document Structure. *Proceedings of the 7th ACM Conference on Multimedia*, 1999, 363-371.

[Irti03] IRTI - Innovative Rehabilitation Technology inc. eClipseReader Home Page, 2003. <http://www.eclipsereader.com/>

[James97] F. James. Presenting HTML Structure in Audio: User Satisfaction with Audio Hypertext. *Proceedings of ICAD'97*, 1997, 97-103.

[Martin98] Martin, J.-C., Julia, L., and Cheyer, A. A Theoretical Framework for Multimodal User Studies. *Proceedings of CMC'98*, 1998, 104–110.

[Microsoft02] Microsoft. HTML+TIME 2.0 reference, 2002. <http://msdn.microsoft.com/workshop/author/behaviors/reference/time2_entry.asp>

[Morley98] S. Morley. Digital Talking Books on a PC: A Usability Evaluation of the Prototype Daisy Playback Software. *Proceedings of ASSETS'98*, 1998, 157–164.

[Niso99] NISO. Document Navigation Features List, 1999. <http://www.loc.gov/nls/z3986/background/navigation.htm>

[Oviatt00] Oviatt, S. L., Cohen, P. R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., and Ferro, D. Designing the User Interface for Multimodal Speech and Gesture Applications: State-of-the-art Systems and Research Directions. *Human Computer Interaction*, 15(4), 2000, 263–322.

[Van Buskirk95] Van Buskirk, R. and LaLomia, M. A Comparison of Speech and Mouse/Keyboard GUI Navigation, *Proceedings of CHI'95*, 1995.