8

Audio Interfaces for Improved Accessibility

Carlos Duarte and Luís Carriço
LaSIGE and Faculty of Sciences, University of Lisbon
Portugal

1. Introduction

According to the World Health Organization the number of people with visual impairments worldwide in 2002 was in excess of 161 million, of whom about 37 million were blind (Resnikoff et al., 2004). Although the visually impaired population is not uniformly distributed over the world, estimates for the developed countries, including the United States of America and European Union countries, go up to more than 20 million visually impaired people. Even if considering only the numbers for the developed countries, there are large numbers of population being prevented to fully access, depending on the severity of their visual impairment, today's software applications, which are mostly based on visual interaction.

The limitations to the visually impaired population caused by this reliance on visual interaction are felt both on the input and output ends of the interaction spectrum. Considering the use of visual output modalities, the limitations can range from total content inaccessibility felt by blind users, to minor limitations that are still detrimental to the user experience. These include small font sizes that make it hard to read, colour selections disregarding the problem of the colour blind population, and other presentation related issues. Input modalities are also extremely reliant on visual interaction. Although even the blind population is capable of using the traditional keyboard, pointing devices, like the mouse, are unusable by people with serious visual limitations, which hinder their perception of the pointer representation on screen.

In order to improve accessibility, alternative modalities must be considered. Audio interaction is the most promising alternative to visual interaction for visually impaired users, as the recommendations toward using screen readers and voice recognition software show (W3C, 2008; Sutton, 2002). It can be used alone for users with severe visual impairments who won't benefit from any kind of visual representation, or it can be used as a complementary or redundant modality for visual interaction, assuming greater or lesser relevance in accordance to the visual impairment level of the user population.

This chapter reflects on how audio interaction can improve interface accessibility, and shows its usefulness by describing the development of an audio based interface for Digital Talking Book (DTB) listening. DTBs are primarily targeted at blind and vision impaired users, but their development under the Universal Accessibility (Stephanidis & Savidis, 2001) umbrella can extend their usage to settings where sighted users operate in constrained environments that restrict visual interaction.

The chapter begins with a short summary of the issues pertaining to DTB presentation. This is relevant since a DTB player will be the application used to illustrate how audio interaction

can increase interface accessibility. This is followed in section 3 by a study comparing the use of auditory icons, earcons and speech in an audio only interface for a DTB player. The different techniques are evaluated according to the identification errors made, and subjective measures of understandability, intrusiveness, and likability. Section 4 presents the recommendations resulting from this study.
These recommendations are then accounted for during the development of a DTB player for two platforms: a desktop and a PDA. Section 5 will present the development of the DTB player for both platforms. The DTB player uses audio feedback to support non visual interaction. Books are recorded in audio streams which are played back to the user. The audio streams are synchronized with the books textual content allowing for visual presentation if required and supporting a set of additional features, which include improved navigation mechanisms and annotations support. Navigation possibilities are offered through the table of contents, and user defined bookmarks. The DTB player interface also supports annotation reading and creation. Audio annotations can be created using the devices' audio recording features. In visual operation, multimedia annotations consisting of images and videos can also be created and visualized.
For visually impaired users, and for visually constraining environments, it is necessary to rely on audio based awareness raising mechanisms. These are integrated into the interface to alert to the presence of navigation elements and existing annotations. The two versions presented in section 5 are visual based, but are introduced to lay the ground for the audio only version presented in section 6, which makes use of the recommendations from section 4 for presentation, and introduces pointer less based interaction, allowing blind users to control the application with a reduced number of keys.
Finally, section 7 presents the conclusions.

2. Digital Talking Book Presentation

Digital recordings of book narrations synchronized with their textual counterpart allow for the development of DTBs, supporting advanced navigation and searching capabilities, with the potential to improve the book reading experience for visually impaired users. By introducing the possibility to present, using different output media, the different elements comprising a book (text, tables, and images) we reach the notion of Rich Digital Book (Carriço et al., 2006). These books, in addition to presenting visually or audibly the book's textual content, also present the other elements, and offer support for creating and reading annotations.
Current DTB players do not explore all the possibilities that the DTB format offers. The more advanced players are executed on PC platforms, and require visual interaction for all but the most basic operations, behaving like screen readers, and defeating the purpose to serve blind users (Duarte & Carriço, 2005).
The DTB format, possessing similarities with HTML, has, nevertheless, some advantages from an application building perspective. The most important one is the complete separation of document structure from presentation. Presentation is completely handled by the player, and absent from the digital book document. Navigation wise, the user should be able to move freely inside the book, and access its content at a fine level of detail. The table of contents should also be navigable. One major difference between a DTB player and a HTML browser is the support offered for annotating content. Mechanisms to prevent the

reader becoming lost inside the book, and to raise awareness to the presence of annotations and other elements, like images, are also needed.
To solve these problems in an audio only environment (speech recognition plus auditory display) we have tested several approaches. Concerning the auditory display, playback of pre-recorded books may be complemented by three other solutions: pre-recorded speech cues, auditory icons (Gaver, 1986) and earcons (Blattner et al., 1989). These solutions are used to convey context information and navigational cues, and their comparison is presented in the following section.

3. A Study of Audio Presentation Techniques in Rich Digital Talking Books

DTBs are capable of presenting their contents either on screen or through speech, recorded or synthesized. Besides the main content presentation, other book elements also have to be presented when working in an audio only environment. This means that the table of contents and the annotations must have an audio representation also. If an annotation is a voice annotation this is straightforward. If it is a text annotation, its content can be reproduced using a speech synthesizer.
However, in an audio only environment, not only content has to be transmitted, but the entire narration context has to be available in an audible format, thus enabling the listener to form an accurate image of the surrounding elements. For this to be possible the listener must be aware of annotations and images present in the book, as well as be able to know what is her/his position in the book whenever desired.
To understand how to better transmit this information to the listener, we evaluated three techniques for improving the listener awareness to the different DTB elements: speech, auditory icons and earcons.
Using speech for transmitting context information is perhaps the easiest of the three approaches, involving just the selection and recording or synthesis of the words to employ. While for certain applications this may not be a trivial task, in the DTB context, where the elements are well identified, it is an uncomplicated one. Speech can also be expected to be the technique where the message meaning is most easily understandable by the listener.
However, the use of speech can have disadvantages also. Since the book's content is being narrated, there will be two audio tracks presenting information in the same manner. If the presented messages are long they can disrupt the reading experience, become too intrusive, or even make it harder to listen to the main content if both tracks are played back simultaneously (Petrie et al., 1998). Furthermore, for voice messages to be understood, the listener must know the language in which the messages are spoken.
Auditory icons have been defined by Gaver (Gaver, 1997) as "*Everyday sounds mapped to computer events by analogy with everyday sound-producing events*". Due to this nature, auditory icons share with voice commands the ease of understanding, if enough care is put into the auditory icons selection, ensuring appropriate and intuitive mappings between the sounds and what they represent in the interface. However, there may be cases where it may be difficult, and even impossible, to find a sound to map to abstract interface events or components (Brewster, 2002). In the DTB domain, certain concepts are abstract enough to make it harder to find an everyday sound to map to, e.g. the beginning of a chapter.
Earcons are "*abstract, synthetic tones that can be used in structured combinations to create auditory messages*" (Brewster, 1994). They can be used in the situations where there are no intuitive sound to represent an interface's event. This gives them the advantage of being able to

represent any event or interaction with the interface. They are based on an abstract mapping between a music-like sound and the interface events, which means that, at least initially, they have to be explicitly learned.
There are four types of earcons (Blattner et al., 1989): one-element, compound, hierarchical and transformational, allowing them to be used in every situation, and even giving them the flexibility to be concatenated, in a process similar to building sentences out of words (Brewster, 1994). Guidelines on how to build earcons are also available (Brewster et al., 1995), identifying timbre, rhythm, pitch and register as sound characteristics that can be used to effectively differentiate one earcon from the others.

3.1 Experimental Setting

In order to understand what solutions are more appropriate for the different DTB elements, and how they can be used, an experiment was set up, evaluating the use of the three different techniques, in a purely audio version of the DTB interface. To better focus on this goal we conducted a Wizard of Oz evaluation, with just the features required for an audio environment.
Four elements, essential for contextual awareness, were the subject of evaluation: beginning of a new chapter, current chapter number, presence of an annotation and presence of an image. A pre-recorded narration of the book "O Senhor Ventura" by a professional narrator was used in the experiment. Four excerpts of the narration, each making use of different audio feedback techniques, were prepared:

1. The first excerpt, six minutes and 39 seconds long, consisted of four chapters. Chapter beginnings, the presence of annotations, and the presence of images were signaled with earcons. The current chapter number was transmitted by speech recordings. When listening to the chapter number, the book's narration was paused.
2. The second excerpt, seven minutes and 38 seconds long, consisted of three chapters. Chapter beginnings were announced by a speech recording of the chapter number. Speech was also used to signal the presence of images and annotations. User requests for chapter numbers were answered with earcons, with no interruption of the book's narration.
3. The third excerpt, six minutes and 36 seconds long, consisted of three chapters. Chapter beginnings were announced with an earcon. Auditory icons signaled the presence of images and annotations. Speech recordings were used to transmit the chapter number, without pausing the book's narration.
4. The fourth and last excerpt, six minutes and three seconds long, consisted of three chapters. All feedback was given using earcons. The chapter number announcements paused the book's narration.

The speech used consisted of pre-recordings of the words "*annotation*", "*image*", and of the chapter numbers. Each chapter number was recorded on its own, meaning that there were no composition of recordings of individual numerals.
Auditory icons were used to signal the presence of images and annotations. For the image signal, the sound a photographic camera shutter closing was employed. The sound's duration was 600 milliseconds. For annotations, the sound of a typewriter was used. This sound's duration was 3 seconds and 50 milliseconds. This last sound was larger than the first because it was expected to be more difficult to recognize.

Earcons were designed to signal chapter beginnings, presence of images and annotations, and chapter numbers. Figure 1 presents the earcons used in the evaluation procedure. To promote ease of identification, each earcon is designed according to the earcon design guidelines (Brewster et al., 1995). Different timbres are employed: chapter beginnings – marimba; images – synth bass; annotations – trombone; numerals 1 to 4 – acoustic piano; numerals 5 to 9 – organ; and numeral 0 – tubular bells. The earcons for chapter numbers were divided into three groups corresponding to numerals 1 to 4, numerals 5 to 9 and the numeral zero. The numerals 1 to 4 are played with the same timbre, each numeral consisting of one more note than the previous, played in an ascendant scale. The numerals 5 to 9 are played with another timbre, following the same principles, but with each note played in a descendant scale. The numeral zero is played with yet another timbre. Numbers above 9 were composed with sequential presentation of the individual numerals (e.g. the number 15 is presented by playing the numeral 1 followed by the numeral 5). The interval used between numerals when composing numbers was 400 milliseconds.

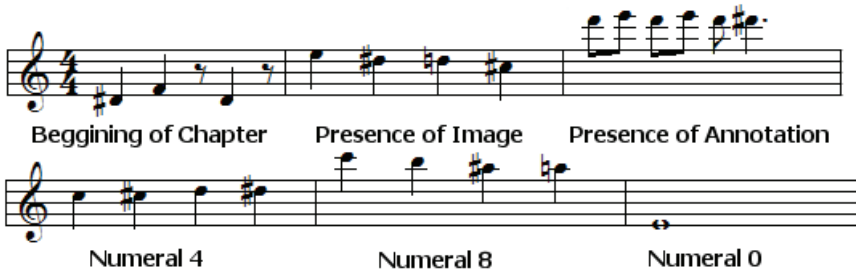

Figure 1. Earcons: Beginning of chapter, presence of image and annotation, and three examples of numeral used for chapter numbers

3.2 Experimental Procedure

Seven participants aged between 21 and 26, one female and six males, undertook the experiment. The experiment was a within-participants factorial design with two independent variables. The first independent variable was the type of auditory feedback technique used. The second variable defined how the current chapter number was presented: with or without interruption of the main narration. Dependent variables were the number of correct identifications of book elements, and subjective measures of understandability, intrusiveness and satisfaction. The main hypothesis was that varying the type of auditory feedback used would lead to different levels of understandability, intrusiveness and satisfaction.

The experiment began with the presentation phase, where the participants were introduced to the different auditory feedback techniques. In this phase the participants were asked to recognize the sounds used as auditory icons, and to associate them with one of the features of the DTB player. This was followed by the presentation of the different earcons, repeated as many times as wished. When the participants felt comfortable with the earcons, these were played back twice in a different order, to test the recall rate. The construction of numbers from the numeral earcons was then explained to the participants. The participants were then asked to identify twelve numbers represented by earcons. This phase ended with the replay of earcons used for beginning of chapter, images and annotations.

The testing phase consisted in the presentation of the four book excerpts. During the excerpts presentation, participants were allowed to use three commands: pause and play, for controlling playback (no forward or backward movement was allowed) and another command to inquire the current reading position. The participants were asked to perform two tasks during excerpt presentation: one task consisted in keeping count of the number of annotations (or images – this varied from excerpt to excerpt) in that excerpt; the second task consisted in identifying, for all occurrences of images (or annotations), the current chapter number, writing it down and delivering it to the test coordinator, immediately after recognizing the audio cue. After each excerpt, the participants answered a questionnaire, rating the techniques used in the excerpt in terms of their understandability, intrusiveness, and satisfaction. Rating scales ranged from zero to nine, with zero meaning it was hard to understand the sound's meaning, the sound was very intrusive, and unpleasant. Nine corresponded to a sound with an easily identifiable meaning, not intrusive, and pleasant to listen to.

3.3 Results

The preparation phase allowed for the individual evaluation of the auditory icons and earcons, while their use as part of an application was evaluated during the next phase.

The auditory icons were correctly identified by all the participants. The sound of the camera shutter closing was associated with the image element by all participants. The typewriter sound was associated with the annotation element by six participants. The other participant associated the sound with the image element.

The three earcons for chapter beginning, images and annotations, presented twice to each participant, were correctly identified by just three participants. One participant was not able to correctly interpret the chapter beginning and annotations earcons in the first round, exchanging their meanings. Two participants exchanged the meanings of the annotations and images earcons in both rounds of presentation. The other participant exchanged the meanings of the beginning of chapter and images earcons in both rounds. No clear misinterpretation pattern was identified. It is possible that all these misinterpretations are due to the participants not having heard the earcons enough times to correctly recall them.

The twelve number earcons for the number identification task represented the numbers 62, 8, 17, 2, 46, 93, 2, 30, 11, 54, 66, 9. Four were single digit numbers, six were composed by earcons of different timbres, and two by earcons of the same timbre. Three participants correctly identified all numbers (the same participants that had correctly identified all the earcons previously). Two participants incorrectly identified two numbers, and the other two participants incorrectly identified five numbers. The fourteen errors, out of the 84 numbers played, can be divided in the following categories: wrong count of notes in a numeral (e.g. identifying a three when a two as played) – 8; wrong association of timbre to numeral (e.g. identifying a two when a six was played) – 3; wrong interpretation of the pause between two notes (e.g. identifying a two when an eleven was played) – 3.The total percentage of correctly identified numbers was 83.3%.

3.3.1 Testing Phase Results

The four excerpts of the book played back to the seven participants contained a total of 385 fixed audio cues divided in the following way: 210 in the form of earcons, 84 in the form of auditory icons and 91 in the form of spoken messages. We will use this number of fixed

audio cues as the corpus for comparison between the different techniques. To arrive at the total number of audio cues, the number of times the chapter number was requested would have to be added.

When considering the identification of audio cues, we expected that both speech and auditory icons would be identified correctly every time. This was indeed the case, with all participants identifying correctly all the elements when presented with these two techniques. When the elements were presented by earcons, the recognition rate lowered to 89.05%., corresponding to a total of 23 misinterpreted earcons over the 7 experiments. The percentages of incorrect interpretations by book element were as follows: 15.71% for the beginning of chapter earcon, 14.29% for the presence of images earcon, and 2.86% for the presence of annotations earcon. This might lead to believe that the beginning of chapter earcon and the presence of images earcons can be misinterpreted one for the other. We can further detail the analysis by looking at how the participants were interpreting the earcons. All the incorrectly identified presence of images earcons were mistaken for presence of annotations, and 72.73% of the beginning of chapter earcons were mistaken for presence of annotations (18.18% were mistaken for numerals and 9.09% for presence of images earcons). These results reveal that the beginning of chapter and the presence of images earcons are not being mistaken one for the other, but are being interpreted as presence of annotations earcons. This is somewhat surprising, since the presence of annotations earcon was correctly identified 97.14% of the times it was played.

The next results report the subjective measures obtained from the questionnaires. The first measure, understandability, can be expected to have a similar outcome to the identification results presented above. Thus, we expected higher values of understandability for speech and auditory icons than for earcons. Figure 2 presents the average understandability for the four elements, by type of audio cue used.

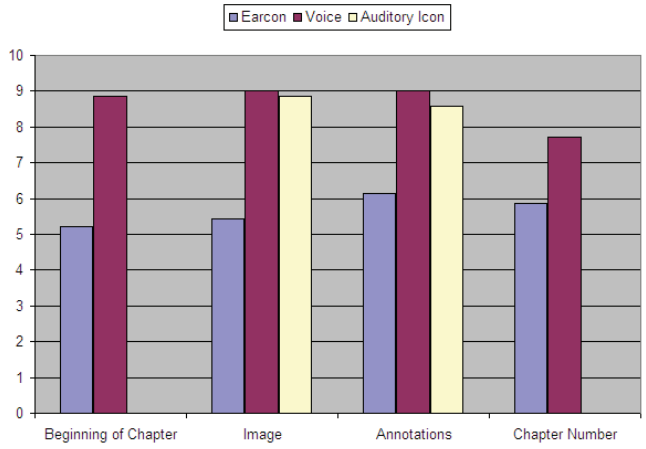

Figure 2. Understandability of the different auditory cues used

To determine if the differences shown are statistically significant two t-tests (one for the beginning of chapter and chapter numbers auditory cues) and two ANOVA tests (one for the presence of images and other for the presence of annotations) were carried out. The t-test comparing the beginning of chapter results found a significant increase ($t(19) = 3.55$, $p <$

0.01) in understandability when speech was used instead of earcons. The t-test comparing the results for chapter numbers between earcons and speech also revealed a significant increase in understandability ($t(26) = 3.03$, $p < 0.01$). The ANOVA for the presence of image cues between earcons, speech and auditory icons was also found to be significant ($F(2, 18) = 40.98$, $p < 0.001$). Post hoc Tukey HSD tests found earcons to have significant lower understandability than speech ($HSD = 11.31$, $p < 0.01$) and auditory icons ($HSD = 10.85$, $p < 0.01$), and no difference between speech and auditory icons. The ANOVA for the presence of annotations understandability when using earcons, speech and auditory icons was also significant ($F(2, 18) = 10.47$, $p < 0.001$). Once again, post hoc Tukey HSD test showed that earcons had significant lower understandability than speech ($HSD = 6.00$, $p < 0.01$) and auditory icons ($HSD = 5.10$, $p < 0.01$). No significant difference was found between speech and auditory icons.

Figure 3 presents the average results for the intrusion rating of the three auditory cues employed (higher values mean less intrusive sounds). Once again, two t-tests and two ANOVA tests were performed to determine if the differences are statistically significant. The t-tests for the beginning of chapter and chapter number comparisons did not identify any significant results. The ANOVA for the intrusiveness when presenting images comparing earcons, speech and auditory icons found a significant difference ($F(2, 18) = 4.01$, $p < 0.05$). Post hoc Tukey HSD tests however did not find significant results between any pair of results. t-tests with the Bonferroni adjustment found that earcons were significantly more intrusive than auditory icons for signaling the presence of an image ($t(12) = 3.62$, $p < 0.05$). The ANOVA test for the presentation of annotations with earcons, speech and auditory icons found a statistically significant difference. The post hoc Tukey HSD tests identified once again that earcons were more intrusive than auditory icons ($HSD = 4.56$, $p < 0.05$).

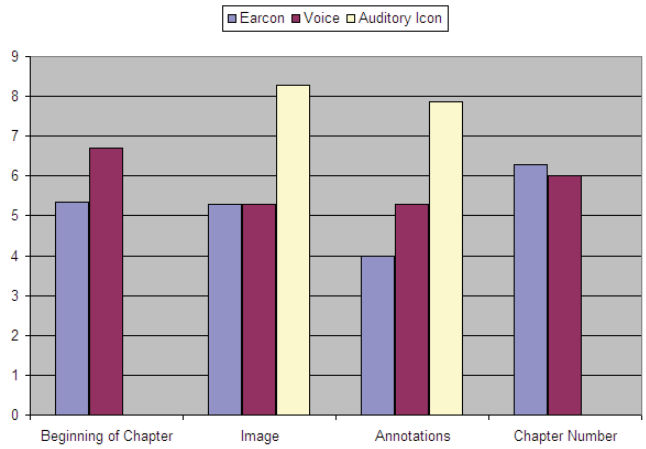

Figure 3. Intrusion of the different auditory cues used. Higher values correspond to less intrusive sounds

Figure 4 presents the average results for the satisfaction rating. The same t-tests and ANOVA tests were applied. The t-test for the beginning of chapter feedback revealed that participants found speech more pleasurable than the earcons ($t(19) = 3.28$, $p < 0.01$). Chapter numbers presented with speech were also found to be significantly more pleasurable than

with earcons ($t(26) = 2.71$, $p < 0.05$). The ANOVA test for the presentation of image presence with earcons, speech and auditory icons found a significant difference ($F(2, 18) = 36.06$, $p < 0.001$). Post hoc Tukey HSD confirms that participants found earcons to be significantly less pleasurable than speech (HSD = 8.65, $p < 0.01$) and auditory icons (HSD = 11.54, $p < 0.01$). The corresponding ANOVA test for annotation presence signaling with earcons, speech and auditory icons also found a significant difference ($F(2, 18) = 13.67$, $p < 0.001$). Post hoc Tukey HSD tests once again confirmed that earcons were found to be significantly less pleasurable than speech (HSD = 5.43, $p < 0.01$) and auditory icons (HSD = 7.06, $p < 0.01$).

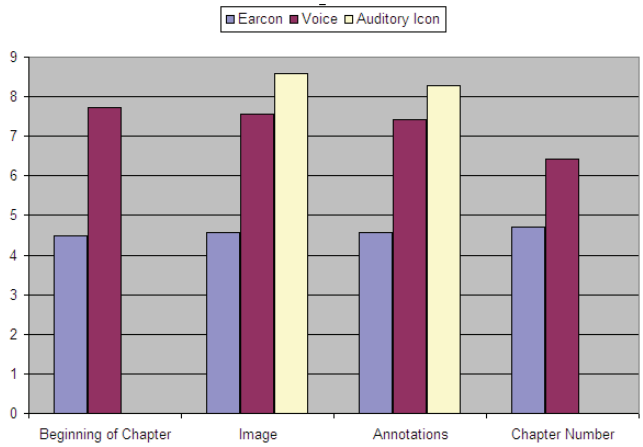

Figure 4. Satisfaction with the different auditory cues used

The effect of interrupting the narration of the main content when presenting the current chapter number on the subjective ratings was also studied. However no significant results were found for understandability, intrusion and satisfaction ratings, which points to the important factor for these ratings being the type of audio feedback technique employed.

4. Audio Design Recommendations

The results presented in the previous section indicate that auditory icons and spoken messages should be preferred to earcons in the design of audio DTB players' interfaces. Earcons proved to be more prone to identification errors, and accordingly, test participants found them less suited to transmit the correct meaning. In addition, the results also show that participants found earcons the least pleasurable of all the evaluated techniques. When considering the intrusion results, earcons and speech achieve comparable results, but both techniques were considered significantly more intrusive than auditory icons.

Observations made during the experiments support these results. It was common amongst test participants to need more time to identify the meaning of a sound when presented with earcons. This is supported by the number of times most participants requested a pause in excerpts which made use of earcons to signal the presence of images or annotations, compared to other excerpts. Another evidence was the request for chapter numbers when they were presented using earcons in comparison with other techniques. Although some participants did the request just for confirmation (the correct number was already written down) it nevertheless shows that participants felt less secure with the earcons.

When comparing test performance on the first and last excerpts, which were the ones which relied most in earcons, all measures evolved positively with the exception of the understandability of the earcon for signaling the presence of an image (average of the answers dropped slightly from 5.43 to 5.29) and the intrusiveness for chapter beginnings, presence of images and annotations. The greatest evolutions were felt in the understandability and likability of the presence of annotations and chapter numbers earcons. This may imply that with more time to familiarize with the earcons used, the measures could continue to evolve positively. However, one cannot be sure until further tests confirm this hypothesis.

For applications sharing the characteristics of a DTB player, we recommend the use of auditory icons and speech. As the events needing audio feedback might not occur frequently in this kind of applications, earcons are at a disadvantage, since it will be harder to memorize and associate their sound with an event, due to the mentioned low frequency of events. The events comprehension should also require the least amount of cognitive effort by the listener, since listening to the book content is the primary task. This is another factor that impacts negatively the use of earcons. We also suggest that auditory icons should be used whenever possible, due to normally being of shorter duration than speech messages. This means smaller interruptions of the book content narration. Another advantage of auditory icons is the fact that they are more universal than any language that may be used, thus requiring less effort for interface development. For the situations where it is difficult to find an auditory icon, then speech can be used to good effect.

5. Visually Enabled Versions of the Rich Book Player

Although ultimately targeting blind users, other visually impaired users can still benefit from the visual component present in the Rich Book Player. Additionally, under Universal Accessibility concerns, the visual component may or may not be used, depending on the context, but the application operation should not be impacted by its presence or absence. As such, the next sections will present two versions of the Rich Book Player, with both visual and audio components. First, a brief overview of a desktop version will be introduced. After, a mobile version will be more thoroughly explained, since that version is the basis for the audio only version.

5.1 The Desktop Rich Book Player

By combining the possibilities offered by multimodal interaction and interface adaptability we have developed the Rich Book Player, an adaptive multimodal Digital Talking Book player (Duarte & Carriço, 2006) for desktop PCs. 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 5 shows the visual interface of the 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 5, 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.

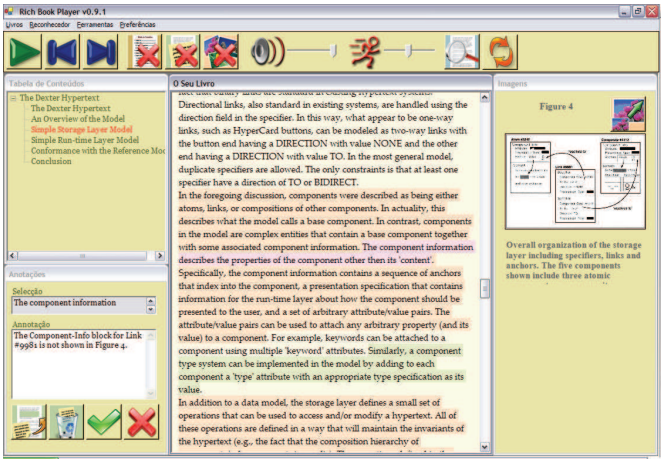

Figure 5. The Rich Book Player's interface. The center window presents the book's main content. On the top left is the table of contents. On the bottom left is the annotations panel. On the right is the figures panel. The content being presented in the player is the article "The Dexter Hypermedia Reference Model", by Halasz and Schwartz

All the visual interaction components have a corresponding audio interaction element, with one exception. Since the speech recognizer currently used in the player 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 speech commands.

5.2 The Mobile Rich Book Player

The mobile version of the Rich Book Player was developed with three main goals in mind: 1) Allow for an anytime, anywhere entertaining and pleasant reading experience; 2) Retain as much as possible of the features available in the desktop version; 3) Support a similar look and feel and foster coherence between both applications.
To achieve these goals, architectural and interaction changes had to be made with regard to the desktop version. The two major limitations of the mobile platform are the limited screen size and processing power.

5.2.1 Main Components Display

Figure 5 displays the main components of the Rich Book Player: main content, table of contents, annotations and images windows. On the desktop version it is possible to display all the components simultaneously and users can find the arrangement that best suits them.

Due to the much smaller screen size of the mobile device it is impossible to follow the same approach.

Figure 6 presents the main content view of the mobile version of the Rich Book Player. The four main areas of interaction can be seen in the figure. On the top, three tabs allow for the selection of the current view. The left tab opens the content view (figure 6, left). The tab header is used to display the current chapter number, which, in this way, is quickly available to the user. The middle tab displays the annotations view. This is the view used to read previously entered annotations and write new ones. The right tab is the images view (figure 6, right). In this tab users can see the images that are part of the book, together with their title and caption. All these contents, book text, annotations and images are displayed in the biggest of the interaction areas. Bellow this area is a toolbar which displays the commands to control book playback, navigation, and other features. The final interaction area is the menu bar located at the bottom of the screen. Besides being used to present the menu, the menu bar is also used to display command buttons whenever necessary.

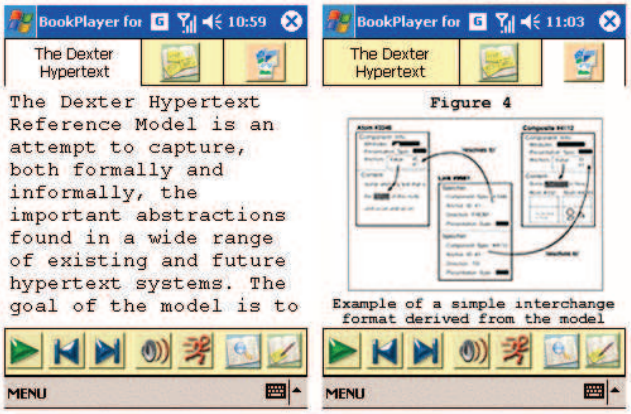

Figure 6. The mobile Rich Book Player. Main content view on the left and images view on the right

Of the four main components of the desktop version, three of them were already mentioned and although they cannot be displayed simultaneously as in the desktop version, they all can be displayed on their own view. The component not yet mentioned is the table of contents. This component has been downgraded to a menu entry and retained just one of its functions. In the desktop version, the table of contents was used to display the current chapter being read, by highlighting its entry, and as a navigation mechanism, allowing users to jump to a particular chapter by selecting its entry. In the mobile version, only the navigation function was retained. The current chapter feedback is now provided as the header of the main content tab.

5.2.2 Annotation Creation and Display

Creating annotations is a two stage process. The user must first select the text to be annotated and only after that enter the annotation. The text selection process is done by selecting the text in the main content view with the stylus and then pressing the create

annotation button (the rightmost button in the toolbar). This takes the user to the annotations view in create annotation mode (figure 7, left).
In the annotations view, the user is still able to see the selected text while entering the annotation. When the text box is selected, the virtual keyboard is displayed, and the *confirm* and *cancel* buttons are reallocated to the menu bar (figure 7, right).

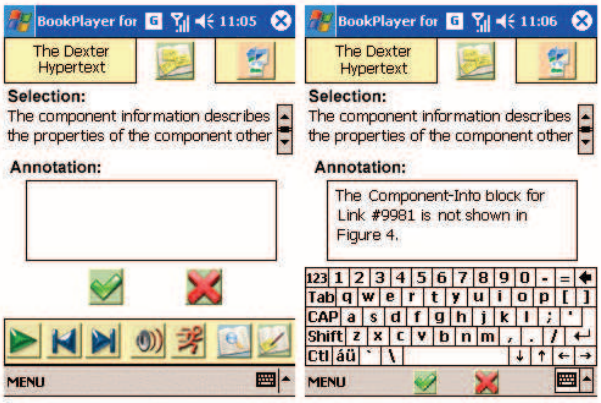

Figure 7. Annotations view during an annotation creation process. On the right, buttons are reallocated to the menu bar when using the virtual keyboard

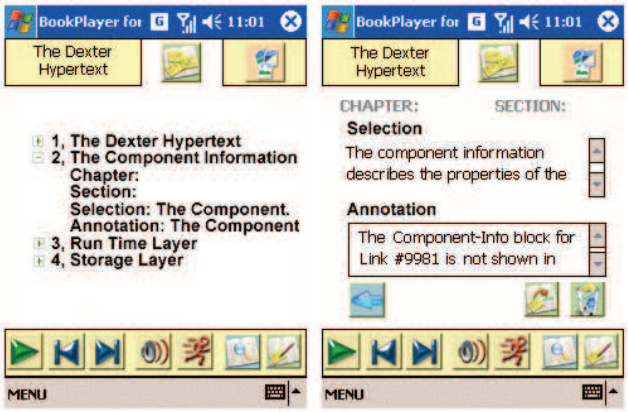

Figure 8. Annotations view in annotation creation mode. Annotations menu on the left, and details of an annotation on the right

Figure 8 presents the annotations view in annotation display mode. When the user changes to the annotations tab, the annotations menu (figure 8, left) displays all the existing annotations in a tree view. By selecting one of the annotations the user is taken to the annotations detail view (figure 8, right). In this view the user can read the current annotation, edit the annotation (which means going to annotation creation mode), delete the annotation, and navigate to the text that has been annotated. The navigation buttons in the toolbar, which in the content view navigate to the next and previous pages, in this view navigate to the next and previous annotations.

5.2.3 Speech Recordings and Synchronization

The main feature distinguishing a Digital Book Player from an e-book player is the possibility to present the book's content using speech, either recorded or synthesized. The desktop version of the Rich Book Player supports both modes of speech presentation. The mobile version currently supports the presentation of recorded speech, using the *Windows Media Player for Pocket PC* for playback.

Speech presentation opens up interaction possibilities that are not available with a visual only interface. With speech, users can change tabs and view images or read annotations while listening to the narration, thus avoiding the forced pause in reading if speech had not been available. Speech also allows users to access the book content without having to look at the device, allowing usage scenarios that, up until now, were available only with portable music devices, but without the limitations of those. For example, performing a search in such a device is extremely cumbersome. In comparison, with the mobile Rich Book Player, search is extremely simple due to the presence of a digital version of the book's text.

With the benefits of incorporating speech into the interface new challenges are uncovered. With speech comes the need for synchronization mechanisms. The application needs these to be able to know when and what images and annotations to present. We were able to port the synchronization mechanism of the desktop version to the mobile version without losing synchronization granularity, meaning the mobile version also supports word synchronization. The synchronization mechanism only had to be adapted to the page concept, introduced in the mobile version, which was absent from the desktop version. The synchronization mechanism allows the application to turn to the next page when the narration reaches the end of the current one. It is also used to visually highlight the word currently being spoken, in order to make the narration easier to follow when the user chooses to both read and listen to the book.

5.2.4 Awareness Raising Mechanisms

Images and annotations require a notification system to alert the user to their presence. Users just listening to the narration need to be alerted through sound signals, while users reading the text need to be alerted through visual signals. Both mechanisms coexist also in the mobile version of the Rich Book Player. Following the recommendations presented in section 4, auditory icons are played when the narration reaches a page with annotations or associated images. Visually, when the user reaches such a page, the annotations or images tabs flash to indicate the presence of an annotation or image.

5.2.5 Pagination

One of the main presentation and interaction differences between the desktop and mobile versions of the Rich Book Player is the introduction of the page concept in the mobile version. The main motivation behind this decision was the desire to avoid the use of scroll bars to read the book. With moderate to large books even small scroll bar movements might give origin to large displacements of the text being displayed, which would quickly turn into a usability problem. Loading a text control with the books full content could also raise performance issues.

To support changing font and font sizes while reading the book, the pagination is executed in real-time. Whenever a book is loaded, or the font settings are altered, a new pagination is started. This implies storing the current reading point (when speech playback is active) or

current page (when speech playback is inactive), repaginate, and present the page of the current reading point. Several choices are possible for the pagination starting point. Starting the pagination algorithm from the book's first page would mean the user would have to wait a variable period for the display to refresh. This period would vary depending on the current reading point. Reading points near the end of book would mean substantially longer waiting periods, due to the lack of processing power of most mobile devices. Starting the pagination from the current reading point, would mean pages with different contents would result from runs of the algorithm with different starting points, even with the same font settings. This might confuse users used to pages holding the same contents on print books. Both approaches raise usability issues. To overcome these issues we employed the notion of forced page break. A forced page break is a location in the text that is guaranteed to be at the start of a page. Employing this notion, the pagination algorithm always produces the same results for the same font settings. The pagination starts from the first forced page break prior to the current reading point and runs to the first forced page break after the current reading point. Since we can control the frequency of forced page breaks, we can guarantee an upper limit on the time that is necessary to paginate until the current reading point, thus assuring the user will not have to wait unacceptably long periods and the pages stay coherent with every run. After this stretch of the book is paginated, the algorithm can run in the background, paginating the rest of the book. Possible choices of forced page breaks, which will be adequate in most situations, are all the entries in the table of contents. This has the added benefit of page breaks being associated with structural book elements, which is something a reader would expect, or, at least, not find confusing.

5.2.6 Adapting the Layout to the Device

Nowadays, mobile devices exist with a multitude of screen resolutions, and even with the possibility of altering the screen orientation. The Rich Book Player visual layout is able to adapt itself to changes in orientation and to different screen resolutions. Figure 9 presents the layout of the Rich Book Player in landscape mode.

Besides changing the layout of the different interface components, a changing in screen orientation also requires a new run of the pagination algorithm, as described in the previous section.

5.2.7 Performance and Storage

The book's presentation involves parallel processing of three threads: the main interaction thread with audio playback, the synchronization thread and the pagination thread. This might impose some performance constraints on the reproduction platform. We have tested the Rich Book Player on two devices: an HP iPAQ h5500 with a 400 MHz XScale processor, 40 MB ROM and 128 MB RAM running Microsoft's Pocket PC 2003, and a QTEK 9100, with a 200 MHz TI OMAP 850 processor, 128 MB ROM and 64 MB RAM running Microsoft's Windows Mobile 5. We have successfully been able to use the application on both devices to read a book of circa 25000 words (a 279 Kb text file), corresponding to a narration with 2 hours and 15 minutes, recorded in a 158 Mb mp3 file.

Due to the size of the audio files, books will have to be made available in storage cards, since it cannot be expected that internal storage of the mobile devices be able to hold such large amounts of data. The Digital Talking Books can be shared between the two platforms, as well as annotations that have been made in one platform can also be read in the other.

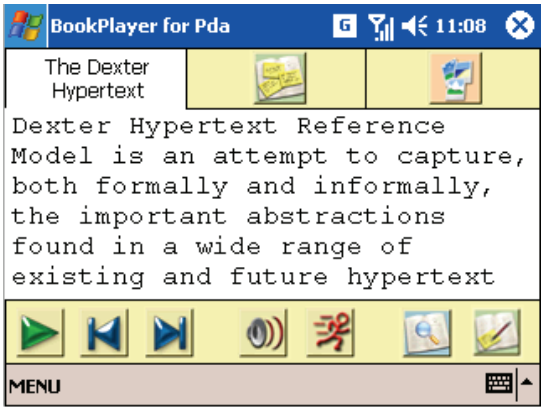

Figure 9. The Rich Book Player layout in landscape mode

6. The Audio Only Version of the Rich Book Player

The audio only version of the Rich Book Player is based on the mobile version described in the previous section. With regard to output capabilities, the mobile version, which incorporates the recommendations from section 4, as described before, is close to fully usable by blind users. This is due to the fact that book contents are recorded in mp3 files, which are played back during book presentation. Additionally, the awareness mechanisms presented are also audio based, like the annotations and image awareness mechanisms which make use of auditory icons. The main issue that is not solved in the player's interface is the presentation of all non-audio annotations. These include text annotations, created using the device's keyboard, and photo or video annotations, captured through the device's camera. Text annotations can be delivered to blind users through speech synthesis when available. Photo or video annotations are currently undeliverable, unless the annotation's author creates an audio file with the annotation description, which could be played back, instead of the default annotation presentation. These limitations do not impact the operation of the Rich Book Player for single use, since a blind user would not write annotations (or use the camera for annotating) when presented with the possibility to speak annotations and have them recorded and latter played back. These limitations are only felt in collaborative scenarios, where the annotations could be shared between readers of the same book. In these scenarios, visually created annotations would be impossible to render using only audio enabled devices.

Concerning the mobile version as described previously, to allow for a completely non-visual interaction, the greatest changes have to impact the input interaction mechanisms. Currently, input is completely visually oriented: button selections, text input, tab selection, menu entries, and text selection. All these tasks rely on stylus operation, which requires the user to be able to look at the screen to know the position of each element to operate. For non-visual operation, an alternative input mean is required. Due to the processing power limitations of current mobile devices, this alternative should not rely on automated speech recognition. It should rely instead on available input capabilities, which can be reliably

employed by blind users. On current mobile devices, the alternative can only be the physical input buttons, which afford recognition by blind users.
It can be safely assumed that all mobile devices (PDAs, mobile phones, with the exception of the iPhone which is not visually impaired friendly) possess a minimal set of physical buttons: a joystick or four directional buttons, with the accompanying selection button, and two more selection buttons. Figure 10 presents this set of buttons in a typical PDA arrangement. Buttons 1, 2 and 7 are selection buttons. Buttons 3, 4, 5 and 6 are directional buttons. Different devices might have a larger number of buttons, but to try to make the application the more device independent as possible, only these seven buttons will be considered from this point forward.

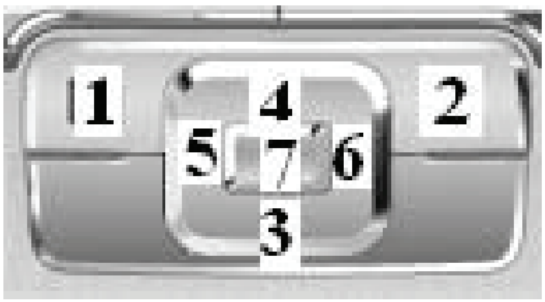

Figure 10. Typical PDA keyboard, with at least 7 different buttons

To enable a fully non-visual operation of the application, all the input commands have to be mapped to these seven buttons. Since there are more than seven commands in the Rich Book Player interface, it will be necessary to map the interface to different states, where each button will have a different meaning. To fully map all the operations, different state diagrams were defined.
Figure 11 presents the first state diagram, representing the key mappings under normal playback. As can be seen in the figure, during playback the directional keys are used to navigate the content. Up and down keys (keys 3 and 4) are used to advance or go back one chapter. Left and right keys (keys 5 and 6) are used to advance or go back one page. The user can pause the playback by pressing key 7. The same key will resume playback when paused. To access the main menu, the user can press the right selection key (key 2). This takes the user to another set of key binding states. During playback, whenever the user wishes to create an annotation, he or she should press the left selection key (key 1). The same key is used to listen to an annotation, whenever the playback reaches a point when there is one available to listen. Annotation creation and annotation listening are also states with different key bindings.
Figure 12 presents the key bindings and state changes when the user is consulting the main menu. As seen before, the main menu is accessed by pressing the right selection key during playback. The same key closes the main menu and takes the user back to the playback mode. In the main menu, the up and down directional keys cycle through all the menu options, with the currently available option being spoken by the interface immediately after the cycling. In this way, the user is aware of what option if available at every instant. This is fundamental when the user is not yet familiar with the menu contents. Later, when the user gets to know the menu contents, he or she does not have to wait for the spoken feedback,

being able to press the selection keys immediately, thus taking advantage of the acquired expertise. If the user selects (using keys 1 or 7) the options "Faster" or "Slower", the application returns to the playback mode, with the playback speed adjusted as per the user's request. If the option selected is "Load" the user will be taken to the book loading mode, and after completing the book selection, the operation resumes in playback mode. The user has two other options available in the main menu. Selecting "TOC" takes the user to the Table of Contents menu. In there, the up and down directional keys cycle through the table of contents entries. In a similar fashion to what happens with the main menu entries, the table of content entries are also spoken by the interface, thus allowing the user to become aware to what chapter or section is currently selected. After selecting one of the entries of the table of contents, the operation is returned to the playback mode, starting from the beginning of the selected table of contents entry. The other option available in the main menu is "Annotations". Selecting this option takes the user to the Annotations menu. This operates in a similar manner to the other menus. Up and down directional keys cycle through the options, and keys 1 or 7 select the annotation which the user desires to listen. After each cycling key press, the annotation identifier is spoken to the user.

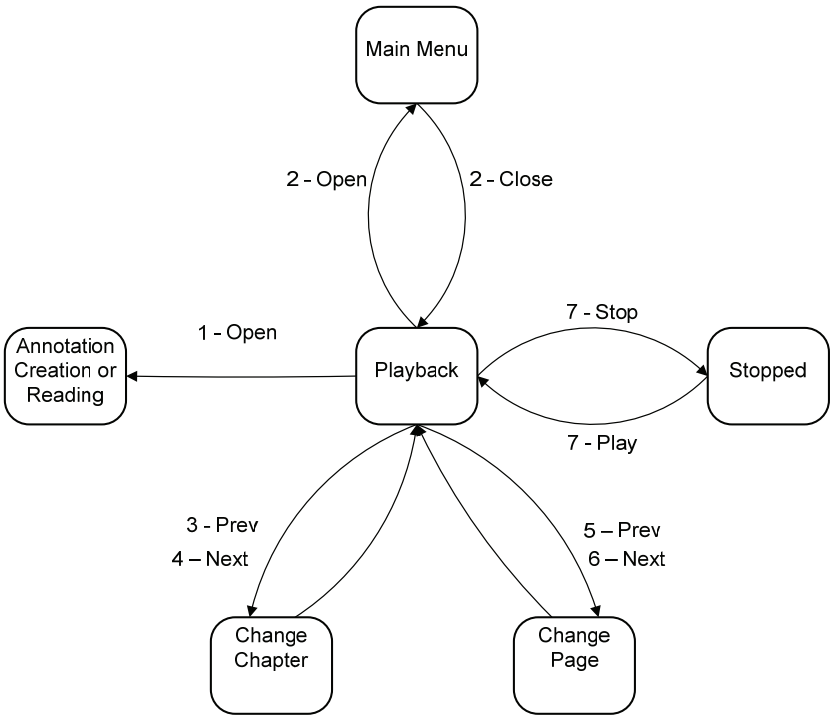

Figure 11. Key mappings and state changes for normal playback conditions

Finally, figure 13 presents the key mappings and state changes for annotation creation and reading operations. In annotation reading mode, left and right directional arrows cycle through available annotations, while the up directional arrow repeats the current

annotation. If the user presses the left selection key (key 1), she will be asked if she wishes to delete the current annotation. This operation can be acknowledge by pressing key 2, or cancelled by pressing key 1. Notice the acknowledgment key is a different key from the one that starts the operation in order to avoid situations where the user presses the same key twice by mistake. If the delete is confirmed the application resumes the book playback. If it is not, the application returns to the Annotations menu. In addition to being able to delete an annotation, the user can also modify it. To achieve this, the key 7 must be pressed. This will take the user to the same operation mode that is called when the user creates an annotation from the playback mode. The user is then asked to utter the required annotation content. When finished the user presses the left selection key. The application then requests confirmation. If the user is satisfied with the recorded annotation she confirms by pressing the right selection key which saves the annotations and returns to playback mode. If not, the operation can be cancelled by pressing the left selection key. The application then asks the user if she wishes to repeat the recording. If the answer is affirmative, the process begins again. If the answer is negative then the operation is cancelled without any annotation being recorded, and playback ensues.

Comparing the audio only version with the other versions of the Rich Book Player, the main lacking feature is image support. However, since the target population for this version are blind users, there is no need to visually display the images. Instead, if there is a recorded description of the image content, it can be processed in exactly the same fashion as an annotation. If there is no recorded description of the image, it can be disregarded, since there is no way to present it to a blind user.

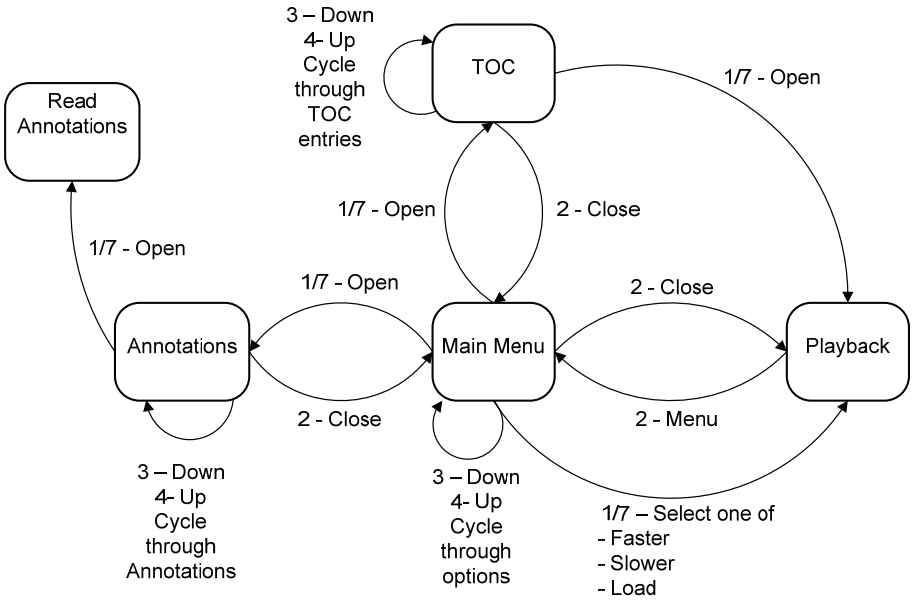

Figure 12. Key mappings and state changes for menu operation

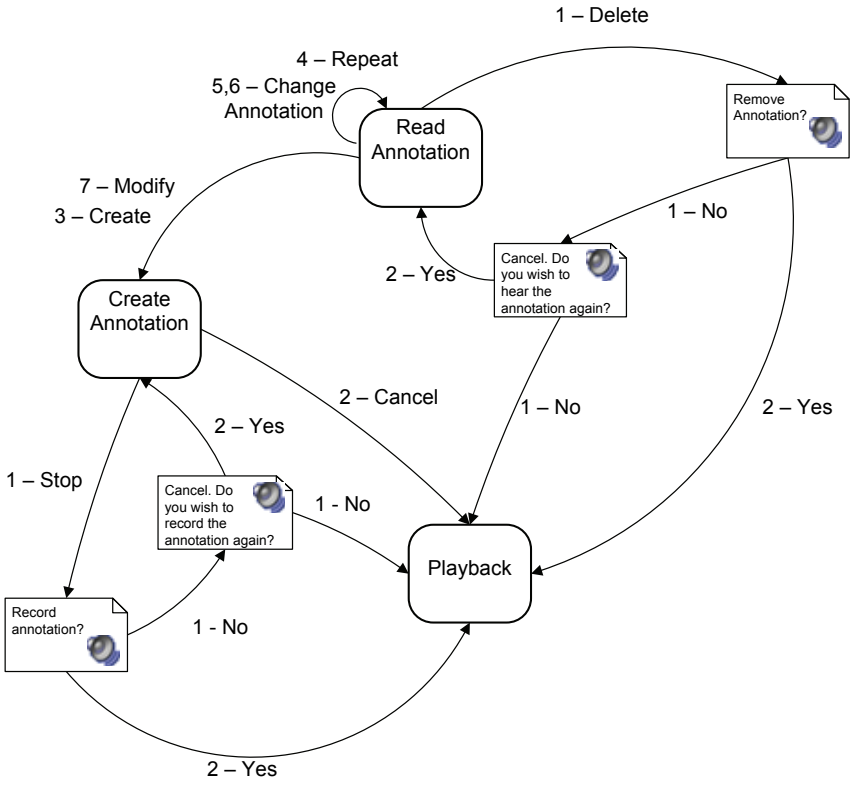

Figure 13. Key mappings and state changes for annotation reading and creation operations

7. Conclusion

This chapter focused on how endowing interfaces with audio interaction capabilities can improve their accessibility. To exemplify this outcome the development of several versions of a Digital Talking Book player was presented. This allowed us to show it is possible to maintain the same set of features while stripping the interface of visual components, and still keep it usable for the visually impaired population.

The interface development concerns focused on both ends of the interaction spectrum: the input and the output. Both these are traditionally very reliant on visual information. To overcome this dependence, visual output was replaced by audio output. On the input side, touch interaction, which is completely based on specific locations on the screen, thus requiring visual inspection, was replaced by mapping all the input options to a minimal set of physical buttons available on the majority of interaction devices, which are able to afford their locations to blind users. This, together with audio feedback, proved capable to convey to blind users the complete interaction features provided by a Digital Talking Book player.

In this chapter we begun by presenting a study of different audio techniques for transmitting awareness information, which compared speech recordings, auditory icons and

earcons. The study results suggest the use of auditory icons combined with speech whenever necessary, in detriment to the use of earcons, for applications with the characteristics of a Digital Talking Book player.

These results were then applied to the development of the Rich Book Player, being essential to increase the usability and accessibility of the mobile version. This same version was used as the foundation of the audio only version of the Rich Book Player. The fundamental difference between the two versions was given by the introduction of the possibility to operate all the commands from the mobile device physical input buttons. In this fashion, all of the application's functionalities are made available through the device's input buttons, complemented with audio feedback, waiving the necessity of using the stylus, and making the application fully accessible to visually impaired users.

This means that a final version, with the full audio interaction capabilities combined with the visual features of the mobile version, can be said to follow the guidelines of Universally Accessible applications. This version makes itself accessible and usable to users with and without visual impairments. Furthermore, non visually impaired users can still use the application in all kinds of settings, even the ones where visual attention needs to be focused elsewhere, by taking advantage of the multiple input and output modalities available.

8. References

Blattner, M.; Sumikawa, D. & Greenberg, R. (1989). Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, Vol. 4, No. 1, pp. 11-44, ISSN 0737-0024

Brewster, S. (1994). *Providing a Strucutred Method for Integrating Non-Speech Audio into Human-Computer Interfaces*. PhD Thesis. Department of Computer Science, University of Glasgow

Brewster, S.; Wright, P. & Edwards, A. (1995). Experimentally derived guidelines for the creation of earcons. *Proceedings of the British Computer Society Human-Computer Interaction Group Annual Conference*, pp. 155-159, Huddersfield, UK, August 1995, Cambridge University Press

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing*, Vol. 6, No. 3, May 2002, pp. 188-205, ISSN 1617-4909

Carriço, L.; Duarte, C.; Lopes, R.; Rodrigues, M. & Guimarães, N. (2005). Building Rich User Interfaces for Digital Talking Books, In: *Computer-Aided Design of User Interfaces IV*, Jacob, R.; Limbourg, Q. & Vanderdonckt, J., (Ed.), 335-348, Springer-Verlag, ISBN: 1-4020-3145-9 (Print) 1-4020-3304-4 (e-book), Berlin Heidelberg

Duarte, C. & Carriço, L. (2005). Users and Usage Driven Adaptation of Digital Talking Books, *Proceedings of the 11th International Conference on Human-Computer Interaction*, Las Vegas, Nevada, USA, July 2005, Lawrence Erlbaum Associates, Inc.

Duarte, C. & Carriço, L. (2006). A Conceptual Framework for Developing Adaptive Multimodal Applications, *Proceedings of the 11th ACM International Conference on Intelligent User Interfaces*, pp. 132-139, Sydney, Australia, January 2006, ACM Press, New York

Gaver, W. (1986). Auditory Icons: Using Sound in Computer Interfaces. *Human-Computer Interaction*, Vol. 2, No. 2, pp. 167-177, ISSN 0737-0024

Gaver, W. (1997). Auditory Interfaces, In: *Handbook of Human-Computer Interaction, 2nd edition*, Helander, M., Landauer, T. & Prabhu, P., (Ed.), pp. 1003-1041, Elsevier, ISBN 0444818766, Amsterdam

Petrie, H.; Johnson, V.; Furner, S. & Strothotte, T. (1998). Design Lifecycles and Wearable Computers for Users with Disabilities. *Proceedings of the First International Workshop of Human Computer Interaction with Mobile Devices*, Glasgow, Scotland, May 1998, Department of Computing Science, University of Glasgow, Glasgow

Resnikoff, S.; Pascolini, D.; Etya'ale, D.; Kocur, I.; Pararajasegaram, R.; Pokharel, G. & Mariotti, S. (2004). Global data on visual impairment in the year 2002. *Bulletin of the World Health Organization*, Vol. 82, No. 11, November 2004, pp. 844-852, ISSN 0042-9686

Stephanidis, C. & Savidis, A. (2001). Universal Access in the Information Society: Methods, Tools, and Interaction Technologies. *Universal Access in the Information Society*, Vol. 1, No. 1, June 2001, pp. 40-55, ISSN 1615-5289

Sutton, J. (2002). *A Guide to Making Documents Accessible to People Who Are Blind or Visually Impaired*, American Council of the Blind, Washington, DC

W3C (2008). Web Accessibility Inititative. Available at http://www.w3.org/WAI/

Advances in Human Computer Interaction

Edited by Shane Pinder

ISBN 978-953-7619-15-2
Hard cover, 600 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008

In these 34 chapters, we survey the broad disciplines that loosely inhabit the study and practice of human-computer interaction. Our authors are passionate advocates of innovative applications, novel approaches, and modern advances in this exciting and developing field. It is our wish that the reader consider not only what our authors have written and the experimentation they have described, but also the examples they have set.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Carlos Duarte and Luís Carrico (2008). Audio Interfaces for Improved Accessibility, Advances in Human Computer Interaction, Shane Pinder (Ed.), ISBN: 978-953-7619-15-2, InTech, Available from: http://www.intechopen.com/books/advances_in_human_computer_interaction/audio_interfaces_for_improved_accessibility

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2008 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.