Conveying Browsing Context Through Audio on Digital Talking Books

Carlos Duarte and Luís Carriço

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

Abstract. This paper presents the results of a study comparing the use of auditory icons, earcons and speech in an audio only interface for a digital talking book player. The different techniques were evaluated according to the identification errors made, and subjective measures of understandability, intrusiveness and pleasurability. Results suggest the use of auditory icons combined with speech whenever necessary, in detriment to the use of earcons, for applications sharing the characteristics of digital talking book players.

Keywords: Evaluation, Audio Interfaces, Auditory Icons, Earcons, Speech, Digital Talking Books.

1 Introduction

Digital recordings of book narrations synchronized with their textual counterpart allow for the development of digital talking books (DTB), 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, images) we reach the notion of Rich Digital Book [1]. 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 offer. 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 [2]. We have developed an adaptive multimodal player, supporting visual and audio interaction, able to adapt its interaction to user characteristics [3]. In this paper we detail the approach taken in the development of the player features that specifically target blind users.

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

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 259–268, 2007.

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) several approaches have been tested. Concerning the auditory display, playback of pre-recorded books is complemented with three other solutions: pre-recorded speech cues, auditory icons [4] and earcons [5]. These solutions are used to convey context information and navigational cues.

This paper presents the evaluation of several approaches to the problem of how to convey the current context while browsing a rich DTB. The following section details the different book presentation elements, and the approaches that have been tested for conveying their presence. Section 3 presents the experimental setting used in the evaluation of the different approaches. The evaluation results are given in section 4, and section 5 discusses these results, and derives a set of guidelines for developing applications with characteristics similar to those of a DTB player.

2 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 annotation must have an audio representation also. If the 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 reader to form an accurate image of the surrounding elements. For this to be possible the reader 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 reader, we evaluated three techniques for improving the reader 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 way. 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 [6]. 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 [7] 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 [8]. 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*" [9]. 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 [5]: 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 [9]. Guidelines on how to build earcons are also available [10], identifying timbre, rhythm, pitch and register as sound characteristics that can be used to effectively differentiate one earcon from the others.

3 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 decided not to use the current version of the DTB player, instead conducting 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 with 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.

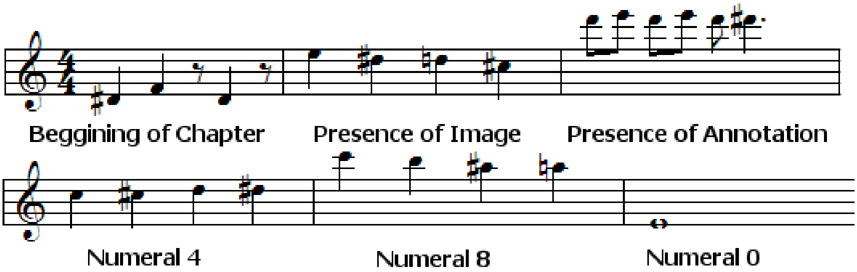

Fig. 1. Earcons: Beginning of chapter, presence of image and annotation, and three examples of numeral used for chapter numbers

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 [10]. 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 a different 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.

3.1 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 pleasurability. 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 pleasurability. 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.

4 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%.

4.1 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 spech 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.

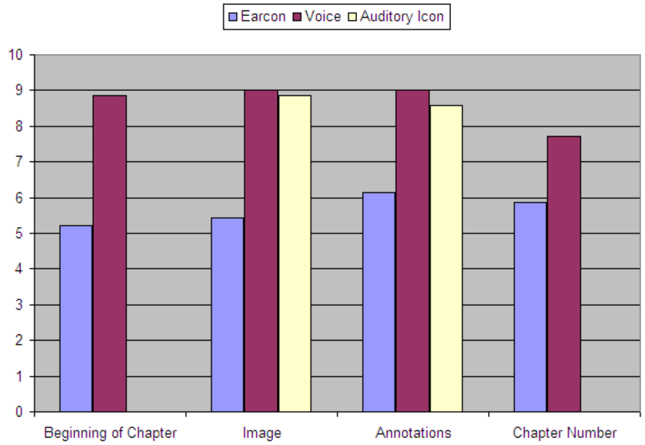

Fig. 2. Understandability of the different auditory cues used

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 be have significant lower understandability than speech (HSD = 11.31, $p < 0.01$) and than 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, $p < 0.05$).

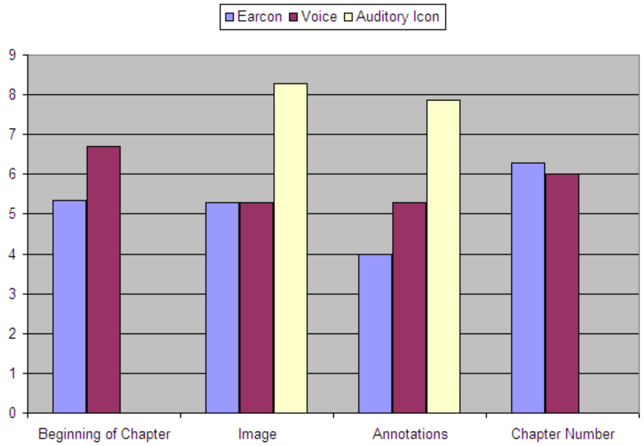

Fig. 3. Intrusion of the different auditory cues used. Higher values correspond to less intrusive sounds.

Figure 4 presents the average results for the pleasurability 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, $p < 0.01$).

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 pleasurability ratings, which points to the important factor for these ratings being the type of audio feedback technique employed.

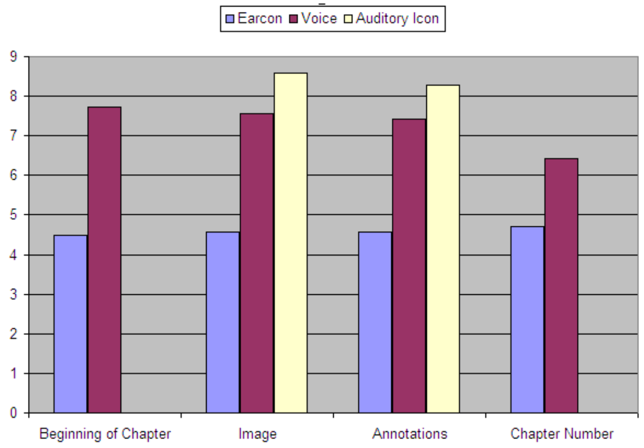

Fig. 4. Pleasurability of the different auditory cues used

5 Conclusions

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 pleasurability 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.

References

1. Carriço, L., Duarte, C., Lopes, R., Rodrigues, M., Guimarães, N.: Building Rich User Interfaces for Digital Talking Books. In: Jacob, R.J.K., Limbourg, Q., Vanderdonckt, J. (eds.) Computer-Aided Design of User Interfaces IV, pp. 335–348. Springer, Heidelberg (2005)
2. Duarte, C., Carriço, L.: Users and Usage Driven Adaptation of Digital Talking Books. In: Proceedings of the 11th International Conference on Human-Computer Interaction, Las Vegas, Nevada, USA (2005)
3. Duarte, C., Carriço, L.: A conceptual framework for developing adaptive multimodal applications. In: Proceedings of the 11th International Conference on Intelligent User Interfaces, Sydney, Australia, pp. 132–139. ACM Press, New York (2006)
4. Gaver, W.W.: Auditory Icons: Using Sound in Computer Interfaces. Human-Computer Interaction, vol. 2(2), pp. 167–177. Lawrence Erlbaum Associates, Inc. Mahwah, NJ (1986)
5. Blattner, M., Sumikawa, D., Greenberg, R.: Earcons and Icons: Their Structure and Common Design Principles. Human-Computer Interaction, vol. 4(1), pp. 11–44. Lawrence Erlbaum Associates, Inc. Mahwah, NJ (1989)
6. Petrie, H., Johnson, V., Furner, S., Strothotte, T.: Design Lifecycles and wearable computers for users with disabilities. In: Proceedings of the First International Workshop of Human Computer Interaction with Mobile Devices, Glasgow, Scotland. Department of Computing Science, University of Glasgow (1998)
7. Gaver, W.W.: Auditory Interfaces. In: Helander, M.G., Landauer, T.K., Prabhu, P.V. (eds.) Handbook of Human-Computer Interaction, 2nd edn. vol. 1, pp. 1003–1041. Elsevier, Amsterdam (1997)
8. Brewster, S.A.: Overcoming the lack of screen space on mobile computers. Personal and Ubiquitous Computing, vol. 6(2), pp. 188–205. Springer, Heidelberg, New York (2002)
9. Brewster, S.A.: Providing a Structured Method for Integrating Non-Speech Audio into Human-Computer Interfaces. PhD Thesis, Department of Computer Science, University of York (1994)
10. Brewster, S.A., Wright, P.C., Edwards, A.D.N.: Experimentally derived guidelines for the creation of earcons. In: Proceedings of BCS-HCI, Huddersfield, UK, pp. 155–159. Springer, Heidelberg, New York (1995)