SPEARCONS: SPEECH-BASED EARCONS IMPROVE NAVIGATION PERFORMANCE IN AUDITORY MENUS

Bruce N. Walker, Amanda Nance, and Jeffrey Lindsay

Sonification Lab, School of Psychology
Georgia Institute of Technology
Atlanta, GA, USA 30332
bruce.walker@psych.gatech.edu

ABSTRACT

With shrinking displays and increasing technology use by visually impaired users, it is important to improve usability with non-GUI interfaces such as menus. Using non-speech sounds called *earcons* or *auditory icons* has been proposed to enhance menu navigation. We compared search time and accuracy of menu navigation using four types of auditory representations: speech only; hierarchical earcons; auditory icons; and a new type called *spearcons*. Spearcons are created by speeding up a spoken phrase until it is not recognized as speech. Using a within-subjects design, participants searched a 5 x 5 menu for target items using each type of audio cue. Spearcons and speech-only both led to faster and more accurate menu navigation than auditory icons and hierarchical earcons. There was a significant practice effect for search time, within each type of auditory cue. These results suggest that spearcons are more effective than previous auditory cues in menu-based interfaces, and may lead to better performance and accuracy, as well as more flexible menu structures.

Author Keywords

Spearcons, earcons, auditory icons, speech interfaces, menu navigation.

1. INTRODUCTION

With visual displays shrinking or disappearing due to mobile and ubiquitous computing applications, and with the increasing use of technology by users who cannot look at or cannot see a traditional visual interface, it is important to identify methods or techniques that can improve usability with non-GUI interfaces. Often, non-visual interfaces are implemented via a menu structure. Much is known about good visual menu design [e.g., 1, 2 Ch. 7], however there are still many open questions when it comes to non-visual menus. The use of non-speech audio cues called *earcons* [3] has been suggested as one way to improve auditory menu-based interfaces in a number of ways. While generally promising, there are shortcomings to the use of earcons (and related audio cues such as *auditory icons* [4]), which may be resolved with the introduction of a novel method of creating auditory cues, which we call *spearcons*. In this paper we discuss the potential benefits of spearcons, and then present an empirical evaluation of their effectiveness compared to earcons, to auditory icons, and to spoken menu items with no added auditory cues. The potential benefit of this new technique will be to improve performance and usability of menu-based interfaces, as well as to make many interfaces more accessible to a broader group of users, in a wider range of applications and situations.

1.1. The Use of Earcons and Auditory Icons

Earcons are brief musical melodies consisting of a few notes whose timbre, register, and tempo are manipulated systematically, to build up a "family of sounds" whose attributes reflect the structure of a hierarchy of information [5]. Using earcons has often been proposed as a method to add context to a menu in a user interface, helping the user maintain awareness of where in the tree he or she is currently located. Such context earcons have been applied to menus ranging from graphical user interface (GUI) applications [6], to mobile phones [7], and telephone-based interfaces [8, 9]. Menus in GUIs may also be improved by adding earcons to help prevent the user from selecting the wrong menu item, or from "slipping off" a chosen item [10]. Additionally, earcons have been proposed as a way to help speed up a speech-based interface, including those designed for visually impaired users [e.g., 11], as well as those intended for general usage such as in-vehicle displays [e.g., 12]. In these applications, the sound is meant to help the user know what the content of a menu item is, not just where it is in the menu hierarchy [see also, 13].

One alternative to earcons are auditory icons [4]. These are generally non-musical sounds that have some resemblance to the thing they are representing. That is, an auditory icon representing a printer might sound like a dot-matrix printer or typewriter. Clearly the level of direct resemblance between the auditory icon and the represented item can vary, just as with a visual icon. At some point, the direct iconic representation gives way to a metaphorical representation [see 14]. It should be said that there seem to be few examples of the addition of earcons or auditory icons leading to significantly better performance with auditory menu-based interfaces, in terms of navigation or menu item identification.

1.2. Issues with Earcons and Auditory Icons

When using either earcons or auditory icons in an interface, there are some important issues relating to the effectiveness of the sounds, the ease of creating and maintaining the interface, and the resulting flexibility of the auditory menu interface.

As discussed, earcons can represent location in a menu, as well as menu item content. This makes them potentially more informative than auditory icons, which are mostly effective at conveying content, rather than hierarchical position. In addition, since they use an arbitrary mapping, virtually any set of concepts (i.e., any menu) can be represented by earcons, whereas auditory icons are difficult to create for many menus, especially those in computer interfaces that have no real sound (e.g., "Connect to Server" or "Export File"). However, the arbitrary mapping of earcons means that more learning, and potentially more explicit training, is required for earcons to be

effective. Further, there is potentially very limited transfer of training when moving between two systems employing different earcon "languages".

From a systems engineering perspective, the menus that use either earcons or auditory icons are *brittle*, in that a change to either the menu hierarchy or menu items is not well supported by the sounds. If a menu or menu item needs to be added, then new sounds need to be generated. The hierarchical earcon approach can handle this automatically, so long as the menu or menu item is added *after* the existing items. For example, adding an item to the bottom of a menu would mean that the next timbre or tempo from a preset list would be used to create the earcon appropriately. This requires that the method for creating earcons anticipates a great enough variety in menu items to handle the menu growth. This can be hard to predict, especially for systems that have varied usage, or long life expectancies. Perhaps more problematic is when a menu item is entered in the middle of a menu. For example, if the first item in a file list starts with "C", it is likely that items will subsequently be inserted ahead of it in the list (i.e., as soon as a file whose name starts with "B" is created). Earcons do not handle this situation very well, nor do they handle the related challenge of re-sorting or re-ordering menus (as is often done in "intelligent" menus that bubble the most commonly selected items towards the top). Either the hierarchical order of the earcons must be rearranged, which diminishes their role in providing context, or else the learned mappings for every earcon below the new menu item will need to be relearned. Auditory icons are more flexible in terms of inserting and re-ordering items, but each new item needs to be created manually (assuming an iconic sound can be found for the new item). This is clearly a problem for dynamic systems. To summarize these issues, Figure 1 presents the dimensions of "Ease of sound creation" and "Flexibility of resulting menu". Neither earcons nor auditory icons rate highly in both dimensions. An optimal solution, then, would be sounds that: (1) can be simply and automatically generated; (2) provide less arbitrary mappings than earcons; (3) cover a wider range of menu content than auditory icons; and (4) be flexible enough to support rearranging, resorting, interposition, and deletion of menu items. If such sounds could also increase the speed and/or accuracy of menu selections, they would be even more useful.

1.3. Spearcons: Speech-based earcons

Spearcons are brief audio cues that can play the same roles as earcons and auditory icons, but in a more effective manner, overall. Spearcons are created automatically by converting the text of a menu item (e.g., "Export File") to speech via text-to-speech (TTS), and then speeding up the resulting audio clip (without changing pitch) to the point that it is no longer comprehensible as speech. Spearcons are unique to the specific menu item, just as with auditory icons, though the uniqueness is acoustic, and not semantic or metaphorical. At the same time, though, the similarities in menu item content cause the spearcons to form families of sounds. For example, the spearcons for "Save", "Save As", and "Save As Web Page" are all unique, including being of different lengths. However, they are acoustically similar at the beginning of the sounds, which allows them to be grouped together (even though they are not comprehensible as any particular words). The different lengths help the listener learn the mappings, and provide a "guide to the ear" while scanning down through a menu, just as the ragged right edge of items in a visual menu aids in visual search.

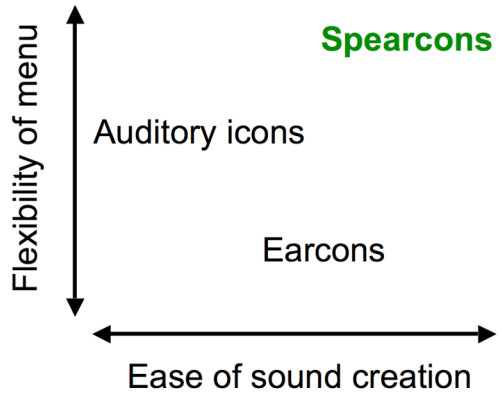

Figure 1. *Relative position of auditory cue types along two axes important in menu effectiveness and usability. In theory, spearcons should be better than previous auditory cue types on both dimensions.*

Since the mapping between spearcons and their menu item is non-arbitrary, there is less learning required than would be the case for a purely arbitrary mapping. The menus resulting from the use of spearcons can be re-arranged, sorted, and have items inserted or deleted, without changing the mapping of the various sounds to menu items. Spearcons can be created algorithmically (though some hand tweaking is sometimes preferable), so they can be created dynamically, and can represent any possible concept. Thus, spearcons should support more "intelligent", flexible, automated, non-brittle menu structures. Now, it should be said that in menus that never change, and where navigation is particularly important (e.g., particularly complex menus), spearcons may not be as effective at communicating their location as hierarchical earcons. However, spearcons would still provide more direct mappings between sound and menu item than earcons, and cover more content domains, more flexibly, than auditory icons. To evaluate this theoretical assessment using real data, we conducted a study comparing menu navigation performance with earcons, auditory icons, and spearcons.

2. METHOD

2.1. Participants

Nine undergraduate students who reported normal or corrected-to-normal hearing and vision participated for partial credit in a psychology course.

2.2. Apparatus and Equipment

A Dell Dimension 4300S PC running Windows XP was used to present the stimuli and collect responses. A software program written in E-Prime [15] was used to run the experiment, including randomization, response collection, and data recording. Listeners wore Sony MDR-7506 headphones, adjusted for fit and comfort.

Animals	Nature	Objects	Instruments	People Sounds
Bird	Wind	Camera	Flute	Sneeze
Dog	Ocean	Typewriter	Trumpet	Cough
Horse	Lightning	Phone	Piano	Laughing
Elephant	Rain	Car	Marimba	Snoring
Cow	Fire	Siren	Violin	Clapping

Table 1. Menu structure used in the present experiment.

2.3. Menu Structure

The menu structure chosen for this experiment is presented in Table 1. In developing this menu, it was important not to bias the study against any of the audio cue methods. For that reason, the menu includes only items for which reasonable auditory icons could be produced. This precluded a computer-like menu (File, Edit, View, etc.), since auditory icons cannot be reliably created for items such as "Select Table". A computer menu was also avoided because that would necessarily be closely tied to a particular kind of interface (e.g., a desktop GUI, or a mobile phone), which would result in confounds relating to previously learned menu orders. This is particularly important in the present study, where it was necessary to be able to re-order the menus and menu items without prior learning causing differential carryover effects. That is, it was important to assess the effectiveness of the sound cues themselves, and not the participant's familiarity with a particular menu hierarchy.

2.4. Auditory Stimuli

2.4.1. Text-to-speech phrases

All of the menu item text labels were converted to speech using Cepstral Text-to-Speech (TTS) [16], except the word "camera", which was produced using AT&T's Text to Speech demo [17]. This exception was made because the Cepstral version of that word was rated as unacceptable during pilot testing. The speech phrases lasted on average 0.57 seconds (range 0.29 – 0.98 sec).

2.4.2. Earcons

For each menu item, hierarchical earcons were created using Apple GarageBand MIDI-based software. On the top level of the menus, the earcons included a continuous tone with varying timbre (instrument), including a pop organ, church bells, and a grand piano; these instruments are built into GarageBand. Each item within a menu used the same continuous tone as its parent. Items within a menu were distinguished by adding different percussion sounds, such as bongo drums or a cymbal crash (also from GarageBand). The earcons lasted on average 1.26 seconds (range 0.31 – 1.67 sec).

2.4.3. Auditory icons

Sounds were identified from sound effects libraries and online resources. The sounds were as directly representative of the menu item as possible. For example, the click of a camera shutter represented "camera"; the neigh of a horse represented "horse". The sounds were manipulated by hand to be brief, and still recognizable. Pilot testing ensured that all of the sounds were generally identifiable as the intended item. The auditory icons averaged 1.37 seconds (range 0.47 – 2.73 sec). Note that for the auditory icon and spearcon conditions, the category titles (e.g., "Animals") were not assigned audio cues—only text-to-speech phrases, as described above.

2.4.4. Spearcons

The TTS phrases were sped up using a pitch-constant time compression to be about 40-50% the length of the original speech sounds. In this study, the spearcons were tweaked by the sound designer to ensure that they were generally not recognizable as speech sounds (though this is not strictly necessary). Thus, spearcons are not simply "fast talking" menu items; they are distinct and unique sounds, albeit acoustically related to the original speech item. They are analogous to a fingerprint—a unique identifier that is only part of the information contained in the original. Spearcons averaged 0.28 seconds (range 0.14 – 0.46 sec).

2.4.5. Combined audio cues and TTS phrases

All of the sounds were converted to WAV files (22.1 kHz, 8 bit), for playback through the E-Prime experiment control program. For three of the listening conditions where there was an auditory cue (earcon, icon, or spearcon) played before the TTS phrase, the audio cue and TTS segment were added together into a single file for ease of manipulation by E-Prime. For example, one file contained the auditory icon for sneeze, plus the TTS phrase "sneeze", separated by a brief silence. This was similar to the approach by Vargas and Anderson [12]. For the "no cue" condition, the TTS phrase was played without any auditory cue in advance, as is typical in many TTS menus, such as in the JAWS screen reader software [18]. The overall sound files averaged 1.66 seconds (range 0.57 – 3.56 sec).

2.5. Task and Conditions

The task was to find specific menu items within the menu hierarchy. On each trial, a target was displayed on the screen, such as, "Find *Dog* on the *Animals* menu." This text appeared on the screen until a target was selected, in order to avoid any effects of a participant's memory for the target item. The menus, themselves, did not have any visual representation—only audio as described above.

The W, A, S, and D keys on the keyboard were used to navigate the menus (e.g., W to go up, A to go left), and the J key was used to select a menu item. When the user moved onto a menu item, the auditory representation (e.g., an earcon followed by the TTS phrase) began to play. Each sound was interruptible such that a participant could navigate to the next

menu item as soon as she recognized that the current one was not the target.

Menus "wrapped," so that navigating "down" a menu from the bottom item would take a participant to the top item in that menu. Moving left or right from a menu title or menu item took the participant to the top of the adjacent menu, as is typical in software menu structures. Once a participant selected an item, visual feedback on the screen indicated whether their selection was correct. Participants were instructed to find the target as quickly as possible while still being accurate. This would be optimized by navigating based on the audio cues whenever possible (i.e., not waiting for the TTS phrase if it was not required). Listeners were also encouraged to avoid passing by the correct item and going back to it. These two instructions were designed to move the listener through the menu as efficiently as possible, pausing only long enough on a menu to determine if it was the target for that trial. On each trial the dependent variables of total time to target and accuracy (correct or incorrect) were recorded. Selecting top-level menu names was possible, but such a selection was considered incorrect even if the selected menu contained the target item.

After each trial in the block, the menus were reordered randomly, and the items within each menu were rearranged randomly, to avoid simple memorization of the location of the menus and items. This was to ensure that listeners were using the sounds to navigate rather than memorizing the menus. This would be typical for new users of a system, or for systems that dynamically rearrange items. The audio cue associated with a given menu item moved with the menu item when it was rearranged. Participants completed 25 trials in a block, locating each menu item once. Each block was repeated twice more for a total of three blocks of the same type of audio cues in a *set* of blocks.

There were four listening conditions: speech only; earcons + speech; auditory icons + speech; and spearcons + speech. Each person performed the task with each type of auditory stimuli for one complete set. This resulted in a total of 4 sets (i.e., 12 blocks, or 300 trials) for each participant. The order of sets in this within-subjects design was counterbalanced using a Latin square.

2.6. Training

At the beginning of each set, participants were taught the meaning of each audio cue that would be used in that condition. During this training period, the speech version of the menu name or item was played, followed by the matching audio cue, followed by the speech version again. These were grouped by menu so that, for example, all animal items were played immediately following the animal menu name. In the speech condition, each menu name or item was simply played twice in a row. Each target item was played once during training.

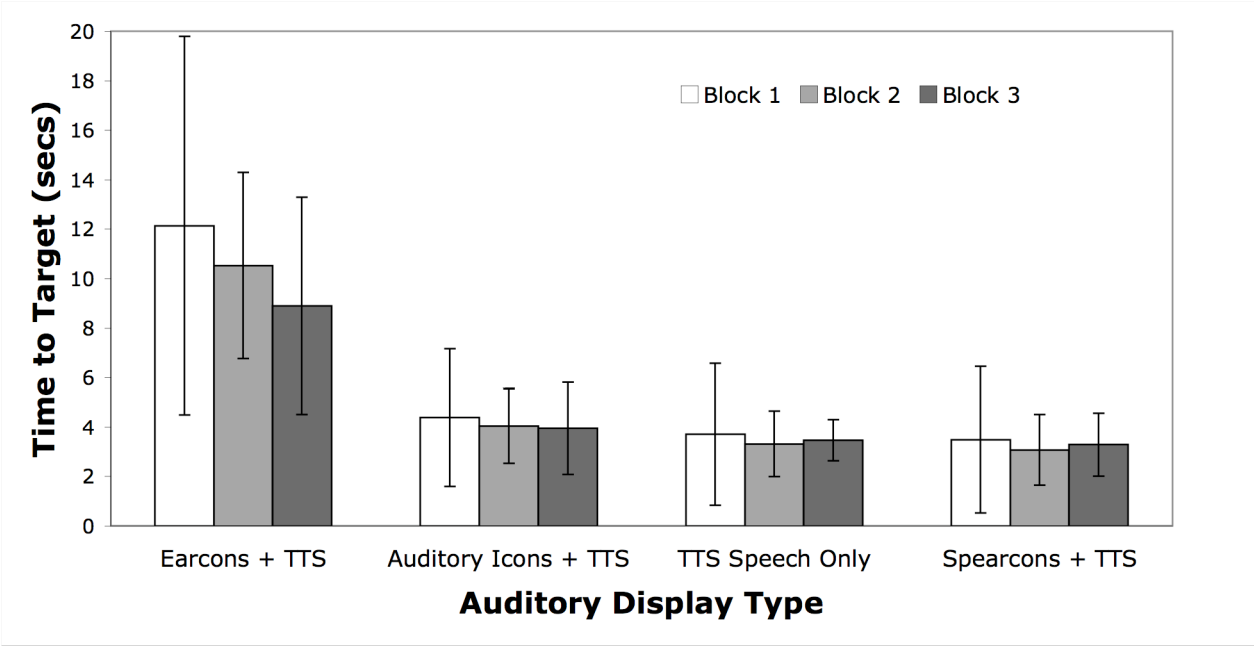

Figure 2. *Mean time to target for each type of auditory display type, for each block within each condition. Note the practice effect, and the relatively poor performance by hierarchical earcons. The speech-only and spearcons+speech conditions were statistically faster than both auditory icons and earcons.*

Type of audio cue	**Mean Time to Target (*SD*) sec.**	**Mean Accuracy (*SD*) %**
Spearcons + TTS phrase	3.28 (.517)	98.1 (1.5)
TTS phrase only	3.49 (.486)	97.6 (2.0)
Auditory icons + TTS phrase	4.12 (.587)	94.7 (3.5)
Hierarchical earcons + TTS phrase	10.52 (11.87)	94.2 (5.4)

Table 2. Overall mean time to target and mean accuracy for each type of audio cue, collapsed across block. Note that spearcons were both faster and more accurate than auditory icons and hierarchical earcons.

3. RESULTS

Figure 2 presents the mean time to target (in seconds) for each audio cue type, split out by the three blocks in each condition. Table 2 summarizes both time to target and accuracy results, collapsing across blocks for simplicity. Considering both time to target and accuracy, together, a multivariate analysis of variance (MANOVA) revealed that there was a significant difference between auditory cue types, $F(3, 6) = 40.20$, $p = .006$, *Wilks' Lambda* = .012, and between trial blocks, $F(5, 4) = 12.92$, $p = .008$, *Wilks' Lambda* = .088.

Univariate tests revealed that time to target (measured in sec.) was significantly different between conditions, $F(3, 24) = 177.14$, $p < .001$, see Table 2. Pairwise comparisons showed that hierarchical earcons were the slowest auditory cue ($p < .001$) followed by auditory icons. Spearcons were faster than the other two cue types ($p = .014$). While spearcons were numerically faster than speech-only (3.28 sec. vs. 3.49 sec., respectively), this difference did not reach statistical significance ($p = .32$) in the present study. Accuracy was significantly different between conditions, $F(3, 24) = 3.73$, $p = .025$, with the same pattern of results (see Table 2) supported statistically.

The practice effect that is evident in Figure 2 is statistically reliable, such that participants generally got faster across the blocks in a condition, $F(2, 24) = 19.17$, $p < .001$. There was no change in accuracy across blocks, $F(2, 24) = 0.14$, $p = .87$, indicating a pure speedup, with no speed-accuracy tradeoff. The fastest earcon block (Block 3) was still much slower than the slowest auditory icon blocks (Block 1; $p = .001$). Anecdotally, a couple of participants noted that using the hierarchical earcons was particularly difficult, even after completing the training and experimental trials.

4. DISCUSSION

Earcons and auditory icons (particularly the former) have been proposed as beneficial additions to auditory menu items. The addition of such audio cues is not typically intended to speed up overall performance (indeed, few, if any authors report performance benefits), but rather to help provide navigational context and help prevent choosing the wrong item, or "slipping off" of the intended item. In the present study, both earcons and auditory icons resulted in slower and less accurate performance than the speech-only condition. This would argue against their usage in a speech-based menu system, at least as far as search performance is concerned. This is not too surprising, since the addition of a 1- or 2-second long sound before each menu item would seem likely to slow down the user. This is particularly true with the earcons, since their hierarchical structure requires a user to listen to most or all of the tune before the exact mapping can be determined. On the other hand, the use of spearcons—speech-based earcons—led to performance that was actually numerically faster and more accurate than speech alone, despite the prepended sound. Spearcons were also clearly faster and more accurate than either earcons or auditory icons. Implementing spearcons in mobile device menus, in telephone-based interfaces for banks and airlines, and in software such as JAWS could lead to a much richer and more effective user experience, with relatively little effort on the part of the developer.

While the performance gains are important on their own, the use of spearcons should also lead to auditory menu structures that are more flexible. Spearcon-enhanced menus can be resorted, and can have items added or deleted dynamically, without disrupting the mappings between sounds and menu items that users will have begun to learn. This supports advanced menu techniques such as bubbling the most frequently chosen item, or the item most likely to be chosen in a given context, to the top of a menu. Such "intelligent" and dynamic menus are not well supported by earcons, and auditory icons are of limited practical utility in modern computing systems where so many concepts have no natural sound associated with them. Spearcons enable interfaces to evolve, as well. That is, new functionality can easily be added, without having to extend the audio design, which increases the life of the product without changing the interface paradigm.

The fact that spearcons are non-arbitrary (which has been discussed here as a benefit), does lead to one possible downside: spearcons are language-dependent, whereas earcons are not. That is, if an interface is translated from, say, English to Spanish, then the spearcons would be different in the two interfaces, whereas an earcon hierarchy would not be different. In some situations this could be problematic. On the other hand, the spearcons can be re-generated automatically, so there is no extra work involved in "internationalizing" an auditory menu with spearcons. Also, Spanish-based spearcons actually sound distinct from English-based spearcons, which is appropriate.

One comment that has been made about spearcons is that perhaps they lead to faster performance simply because they are shorter than earcons and auditory icons. This is probably partially true, but that is simply a structural benefit of spearcons. The musical structure of earcons, and the acoustic realities of auditory icons, essentially "forces" them to be longer, so spearcons have an advantage from the outset, which is reflected in the performance results here. On the other hand, performance is not dependent only on the length of the auditory cue, since auditory icons in this study were longer, on average, than earcons, yet they led to considerably better performance. In any case, none of this discussion about the length of the sounds tarnishes the fact that spearcons also lead to improvements in accuracy.

5. CONCLUSION

As auditory menu-based interfaces become more important and more common, it is important to improve their usability, effectiveness, speed, and accuracy. Spearcons—brief speech-based audio cues—have been shown here to provide all of these benefits, and to do so significantly better than either earcons or auditory icons. In fact, adding spearcons leads to better performance than with the plain text-to-speech menu. In addition, the use of spearcons should allow modern menu interfaces to remain "intelligent," while still incorporating audio cues that are as flexible and dynamic as the interface itself. Spearcons enhance both the system effectiveness, and the user's interaction with the system, which is an important joint outcome in the field of human-computer interaction, especially in novel, and less-well studied interfaces such as audio menus.

6. ACKNOWLEDGMENTS

We would like to thank Yoko Nakano for her help in the later stages of this project.

7. REFERENCES

[1] K. L. Norman, *The psychology of menu selection: Designing cognitive control of the human/computer interface*. Norwood, NJ: Ablex Publishing Corp., 1991.

[2] B. Shneiderman, *Designing the user interface: strategies for effective human-computer-interaction*, 3rd ed. Reading, MA: Addison Wesley Longman, 1998.

[3] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, pp. 11-44, 1989.

[4] W. W. Gaver, "Using and creating auditory icons," in *Auditory display: sonification, audification, and auditory interfaces*, G. Kramer, Ed. Reading, MA: Addison-Wesley, 1994, pp. 417-446.

[5] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, "An evaluation of earcons for use in auditory human-computer interfaces," presented at SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, 1993.

[6] S. Brewster, V.-P. Raty, and A. Kortekangas, "Earcons as a method of providing navigational cues in a menu hierarchy.," presented at HCI'96, Imperial College, London, UK, 1996.

[7] G. LePlâtre and S. Brewster, "Designing non-speech sounds to support navigation in mobile phone menus," presented at International Conference on Auditory Display (ICAD2000), Atlanta, USA, 1998.

[8] S. Brewster, "Navigating telephone-based interfaces with earcons," presented at BCS HCI'97, Bristol, UK, 1997.

[9] S. Brewster, "Using non-speech sounds to provide navigation cues," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 5, pp. 224-259, 1998.

[10] S. Brewster and M. G. Crease, "Correcting menu usability problems with sound," *Behaviour and Information Technology*, vol. 18, pp. 165-177, 1999.

[11] A. Karshmer, P. Brawner, and G. Reiswig, "An experimental sound-based hierarchical menu navigation system for visually handicapped use of graphical user interfaces," presented at ACM ASSETS'94, 1994.

[12] M. L. M. Vargas and S. Anderson, "Combining speech and earcons to assist menu navigation," presented at International Conference on Auditory Display (ICAD2003), Boston, USA, 2003.

[13] C. Wolf, L. Koved, and E. Kunzinger, "Ubiquitous Mail: Speech and graphical interfaces to an integrated voice/email mailbox.," presented at IFIP Interact'95, Lillehammer, Norway, 1995.

[14] B. N. Walker and G. Kramer, "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making," in *Ecological psychoacoustics*, J. G. Neuhoff, Ed. New York: Academic Press, 2004, pp. 150-175.

[15] Psychological Software Tools, "E-Prime, http://www.pstnet.com."

[16] Cepstral Corp., "Cepstral Text-to-Speech, http://www.cepstral.com."

[17] AT&T Research Labs, "AT&T Text-to-Speech Demo, http://www.research.att.com/projects/tts/demo.html."

[18] Freedom Scientific, "JAWS for Windows."