Spearcon Performance and Preference for Auditory Menus on a Mobile Phone

Bruce N. Walker and Anya Kogan

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

Abstract. This study investigates the use of spearcons as an auditory cue. It looks simultaneously at both performance and subjective preference of spearcons and text-to-speech (TTS). The study replicated on a mobile phone a previous PC-based study run by Palladino and Walker [1]. Performance results have been very similar to those found in the previous study, supporting the generalizability of spearcon performance from PCs to mobile phones. TTS and spearcons both provided comparable performance improvements, suggesting that spearcons do not negatively effect the design of visual and non-visual menus and may, within the right context, lead to enhanced designs. Participants gave positive performance scores to both TTS and spearcons when no visual cues were provided. Higher rankings were provided for all audio cues when Spearcons were included both in visual and non-visual conditions.

Keywords: sonification, spearcons, auditory interfaces, auditory menus.

1 Introduction

Many types of auditory displays, and in particular, auditory menus, have been studied either as enhancements to visual displays or as the primary means for interacting with a system or device. Such auditory displays can improve a variety of products, from those with small screens to those being used in limited or no-vision contexts. This may include the use of mobile phones while driving or while walking outside where glare is prevalent. Users with vision impairments can also benefit [2], as many recent GUI designs rely strictly on visual interaction. However, there remain unanswered questions regarding the best ways to design auditory menus.

While most auditory menus are based on simply speaking the menu items to the user (often via text-to-speech, or TTS), this basic approach is now regarded as somewhat simplistic. Many auditory menu enhancement approaches have been considered, in order to maximize the functionality of this new type of interaction. There are several solutions that have been explored most recently as part of auditory menu design. Four approaches that have had considerable attention include: regular speech with no enhancements; adding auditory icons to a speech-based menu [3]; adding earcons [4]; and, as is demonstrated in this study, adding spearcons [5,6]. All

C. Stephanidis (Ed.): Universal Access in HCI, Part II, HCII 2009, LNCS 5615, pp. 445–454, 2009.

of these design approaches have their advantages and disadvantages, many of which are still being studied in order to determine the most proper usage for each.

1.1 Auditory Menus

Increasing the usability and accessibility of menus on small electronic devices is essential due to their decreasing sizes and increasing proliferation. Advanced auditory menus are being studied as an enhancement to the visual-only menus currently on most of these devices, especially when the user is unable to look at the device (e.g., it is in a pocket) or unable to see it (e.g., due to a vision impairment). It remains to be determined how to design an optimal auditory menu, but various enhancements have been proposed to improve the basic (and often unsatisfactory) text-to-speech (TTS) menus often deployed. The study presented here focuses on the use of *spearcons* within auditory menus, but we also explain other approaches, for historical context.

Using sound to enhance menus on electronic devices and desktop computers has the potential to significantly widen the user base. However, most audio menus today are limited to a simple system consisting of a text-to-speech (TTS) conversion of words and phrases. Users can listen to the text provided, use a combination of arrow keys to navigate the menu, and use a select button to choose menu items.

Auditory enhancements have sometimes been prepended to the TTS. The goal of these enhancements is to elevate the efficiency of menu navigation by allowing users to listen to just the cue, if the TTS phrase is not needed for menu navigation. In fact, after becoming familiar with such a system, some users can even turn off the TTS completely and use just the extra audio cues for maximum speed and efficiency.

The transient nature of sound produces several usability challenges to its use in menu design. First, the speech comprehension speed among individuals is highly varied. One study found that blind listeners can understand speech at speeds up to 2.8 times faster than the standard TTS [7]. These differences can be a challenge in creating universal audio cues. Another challenge is location awareness. Users must be able to quickly grasp their location within a menu hierarchy in order to choose the correct path to their desired item [8]. Unlike a visual menu in which users can scan quickly in order to determine their current location, audio menus can take considerable time to present items, and thus can tax the user's working. Further, learning novel auditory cues can be a strain on the user's time, and can lead to poor acceptance; therefore, choosing cues with short learning curves is essential.

Because sound is currently rarely utilized for navigation of menus, there is little information on users' general acceptance of audio cues. It is important to begin to assess user opinions and preferences, since usability depends not only on performance (e.g., time to target), but also subjective impressions. This study will open this topic for research, especially related to the use of auditory menus enhanced with spearcons.

1.2 Auditory Icons

Although this experiment focuses on the use of spearcons, it is important to provide context for their research and development in light of previously developed audio cues. Auditory icons [3] and earcons [4] have been the most popular predecessors to spearcons and will be discussed here. Both have had their advantages and disadvantages, which have been partially addressed with the use of spearcons.

An auditory icon is a direct or metaphorical representation of a word or concept [3], often utilising the sound that the associated word would be known for. For example, a "dog" could be represented with a "woof" sound, and a "cow" with a "moo" sound. For words with clear sound associations, learning can be quick and easy. However, when dealing with electronic menus, clear associations can be difficult or even impossible to create. For instance – what would "delete file" sound like? For this reason, these icons are of limited utility in menu design for modern electronic devices and systems.

1.3 Earcons

Earcons [4] are brief musical motifs or melodies that are used to represent a menu item. Earcons do not require the same natural associations as auditory icons do, and thus can be applied to menus containing any type of information. They can be produced using any systematic set of musical elements that can vary according to frequency, timbre or tempo in order to indicate hierarchy. Guidelines for their design have been suggested by Hereford and Winn [9].

Earcons can present problems due to their rigidity in being able to add or subtract menu items as needed. For example, if an item is inserted into a menu (e.g., adding a new name in a Contact List), the new item would get an earcon assigned to it. However, it is difficult to determine whether it would make sense to keep the earcon assigned to that point in the menu, and move all the other menu items down to be re-assigned to the existing earcons, or else also insert a new earcon for that new menu item. Unfortunately, earcon hierarchies are often fixed, and are based on a musical scale, so inserting a new earcon is generally not possible. Clearly this makes the menu somewhat inflexible, as well. In any case, learning arbitrary earcon-menu item assignments can also be frustrating [6] and difficult for users, even if the mappings do not change. As Walker et al. [5] have stated, the arbitrary nature of the earcons is considered both its strength and its weakness. At the same time, Palladino and Walker [6] showed that listeners learn to associate menu items to spearcons faster than to other types of sounds, such as earcons.

1.4 Spearcons

A spearcon [5], the auditory menu enhancement cue explored in this study, is created by speeding up a spoken phrase without modifying the perceived pitch of the sound. Some of the speech used is compressed so that it is no longer comprehensible as a particular word but a mere representation of that word or phrase, similar to how we think of an icon as a particular image that represents an idea. Walker et al [5] compared the spearcon to a fingerprint because each unique word or phrase creates a unique sound when compressed that distinguishes it from other spearcons. A short learning session leads to easy association of the spearcon to its related word.

Once a spearcon is created, it is prepended to the original spoken word (created either by a TTS generator or a recorded voice) to make a complete, enhanced menu item. A 250 ms pause is typically inserted between the spearcon and the spoken word or phrase. Spearcons are convenient in part due to their brief duration and easy production. More on the production of spearcons can be found in Palladino and Walker's 2008 [1] spearcon study.

There are many advantages to the use of spearcons in auditory menu design. Despite not having natural hierarchical associations, like earcons, Walker et al. [5] found them to result in significantly more efficient navigation than hierarchical earcons [1]. It would also be possible to create additional hierarchical information for the user by augmenting the spearcon with additional audio information if needed.

However, the utility of spearcons in real mobile applications remains to be studied. Desktop applications and mobile phone simulators can provide great insights, but the use of spearcons to enhance menus on a mobile phone may lead to different results. Thus, the present study investigates TTS menus with or without spearcons, and also extends this research paradigm to include an assessment of subjective opinions of these various menu designs.

2 Method

2.1 Participants

A total of 89 undergraduates (55 women and 34 men, mean age = 20) with normal or corrected-to-normal hearing and vision participated for extra credit in psychology courses. English was the native language for 76 of the participants. There were between 15 and 20 participants in each condition.

2.2 Design

This experiment consisted of a between-subjects design with two independent variables. The first was sonification type (TTS Only, Spearcon Cue + TTS, or No Audio), and the second was the visual cues (visual menus were either on or off). Since it would not be feasible to have both no audio and no visual, that condition was not used for this study, leaving five valid experimental conditions.

There were two dependent variables used. The first was the time taken to select the target menu items for each trial. The second was a set of subjective preference scores given to each of the auditory cues used—TTS and Spearcons—individually.

2.3 Materials

Participants used a Nokia N95 mobile phone with a simulated contact list running in Java on the Symbian S60 platform. They used the arrow keys to navigate to desired menu items. They listened to the audio cues through Sony MDR-7506 Dynamic Stereo Headphones. The names used in the contact list (e.g., “Allegra Seidner”) were taken from the study by Palladino and Walker [1], and were produced from a random name generator and translated into sound using the AT&T Labs, Inc TTS Demo program. Spearcons were produced by speeding up the TTS phrases to be very short sounds. The speed-up is logarithmic, so long phrases see a greater compression. The pitch of the sounds is maintained, and the spearcons are still clearly “related to” the original source TTS sounds. More details on spearcon generation can be found in Palladino and Walker’s publication [1].

Each Spearcon + TTS stimuli was created by using Audacity software to prepend the spearcon cue to the TTS with a 250 ms post-cue interval between them. The target name was visibly listed at the top of the phone screen for both the visual on and visual off conditions. In the visuals on conditions, a scrollable list of 50 names were presented, nine of which were visible at a time. A photo of the screen presented can be seen in Figure 1. All 50 names were displayed in alphabetical order by first name. The list scrolled upward or downward according to the key presses. In the visuals off conditions, the list portion of the screen was left blank (i.e., below the target name), though the underlying list was still active and navigable. For all conditions, the list of names did not wrap at the top or bottom of the list to allow for a representative time to target measurement. As each name was placed in focus, both audio and visual cues were presented simultaneously.

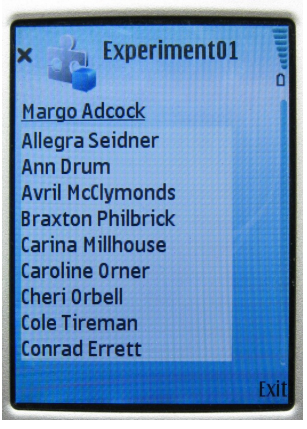

Fig. 1. Screen presented in the condition with the visuals on. The target name, shown underlined at the top of the list, was randomized for each trial.

2.4 Procedure

Participants were assigned to one of five conditions: (1) TTS prepended with a spearcon and no visuals cues; (2) a single TTS cue with no visuals; (3) a single TTS cue with visuals cues; (4) only visual cues with no sound; and (5) TTS prepended with a spearcon and visual cues. Every 25 trials were grouped into a single block, for a total of 10 blocks. Each block was counterbalanced so that one half of the names was used as targets in some blocks and the other half was used in the other blocks. Each block was otherwise identical to all others for a given participant.

Participants were first read aloud a set of instructions that taught about the structure of the menus presented and the required task which was to find the requested target names as quickly and accurately as possible. They were told that they would be timed during the study. Once the participants were given a phone, they could begin by pressing a “continue” key. The timer started once the first down key was pressed. Participants used the up and down arrow keys to reach the target name within the list. Once a name was selected, the end time was recorded and the participant saw the next trial screen with a new target name. Every 25 trials, the participants saw a screen indicating the end of a block and the start of a new one. Each of the nine subsequent blocks proceeded in the exact same way.

After the tenth block, participants filled out a questionnaire that included demographics (i.e. age, gender, native language) and Likert agreement statements to assess their preferences for the TTS and Spearcon audio cues individually. They were only asked to provide their opinions on the cues that were present in their given condition. The scales probed helpfulness, distraction level, preference over silence, fun and annoyance level. A free-response box was also provided for extra comments.

3 Results

3.1 Time to Target

An alpha level of .05 was used for all statistical analysis. Trials with incorrect item selection were disqualified (0.79% of trials in all, 64 in Visuals Off/Spearcons+TTS condition, 21 in Visuals Off/TTS condition, 32 in Visuals On/No Sound condition, 23 in Visuals On/TTS condition, and 38 in Visuals On/Spearcons+TTS condition); a total of 22,072 trial records remained for data analysis. Figure 2 presents the results, specifically the mean time to target for each condition in each block of the experiment. A planned Tukey honestly significant difference was performed on the data to check for significant differences among the different experimental conditions. As expected, overall performance on all conditions including visual cues was significantly faster than those including only auditory cues.

A Tukey honestly significant difference analysis of Block 10 data for each condition found no significant difference between any of the three visuals-on conditions ($p > 0.05$). By Block 10, the significance of the differences in means between the Visuals On/TTS ($M = 6546$, $SD = 3064$) and Visuals On/Spearcons + TTS ($M = 7061$, $SD = 3408$) conditions in Block 10 was very small. It is also clear from Figure 2, that even though the differences between the conditions using auditory-only and auditory with visual cues in Block 10 are significant, there is much less of a difference between the auditory only and visual conditions than existed in the first block of the experiment. Figure 3 illustrates the mean time to target for the five categories in the first and tenth blocks.

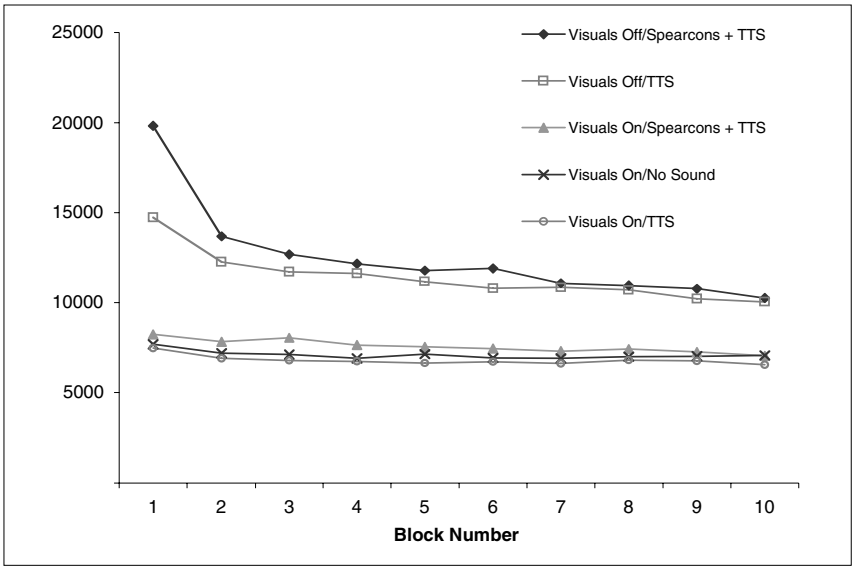

Fig. 2. Mean time to target in milliseconds for all conditions over all blocks

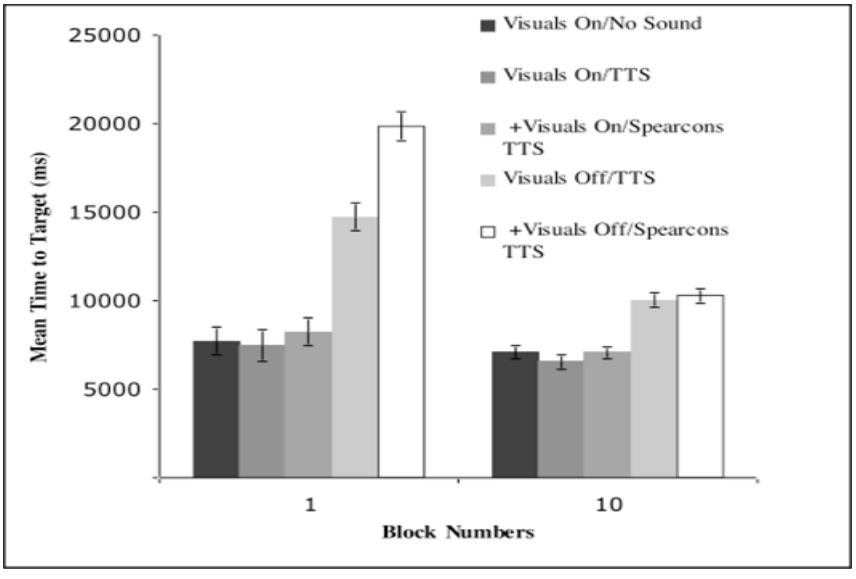

Fig. 3. Mean time to target in milliseconds for all conditions in Blocks 1 & 10. Error bars are 95% confidence intervals.

Collapsing across audio cue types, conditions with the visuals on were significantly faster than visuals off, in both Block 1,$F(1, 2197) = 661.269$, $p < 0.001$, and Block 10, $F(1, 2197) = 348.079$, $p < 0.001$. Considering the different audio cue types (TTS vs. spearcon+TTS), the spearcons cues led to slower times in Block 1, $F(1,2197) = 9.539$, $p = 0.002$, but the effect diminished quickly over the first few blocks, and no significant difference was found amongst the sound conditions for Block 10 ($p > .05$).

3.2 Subjective Ratings

The participants gave scores on five dimensions (i.e. helpfulness, distraction level, preference over silence, fun and annoyance level) by providing agreement or disagreement responses on a Likert scale. The scores were also aggregated into an overall preference score for each participant. The means across all participants for each condition and audio cue are summarized in Figure 4.

Overall, there was no significant difference in preference for spearcons and TTS, $F(1,18) = 3.319$, $p = 0.071$. However, a t-test comparing visuals on and visuals off conditions demonstrated that both audio cues were significantly better rated when no visuals were provided, $t(106) = 6.706$, $p < 0.001$.

The TTS sounds were given significantly higher rankings when they were accompanied by spearcons than when they were not, in both the visuals on condition, $t(33) = -2.234$, $p = 0.032$, and visuals off condition, $t(33)= -3.181$, $p = 0.004$. That is, simply adding spearcons seemed to lead to higher ratings of the TTS, with no performance difference after a few blocks of practice.

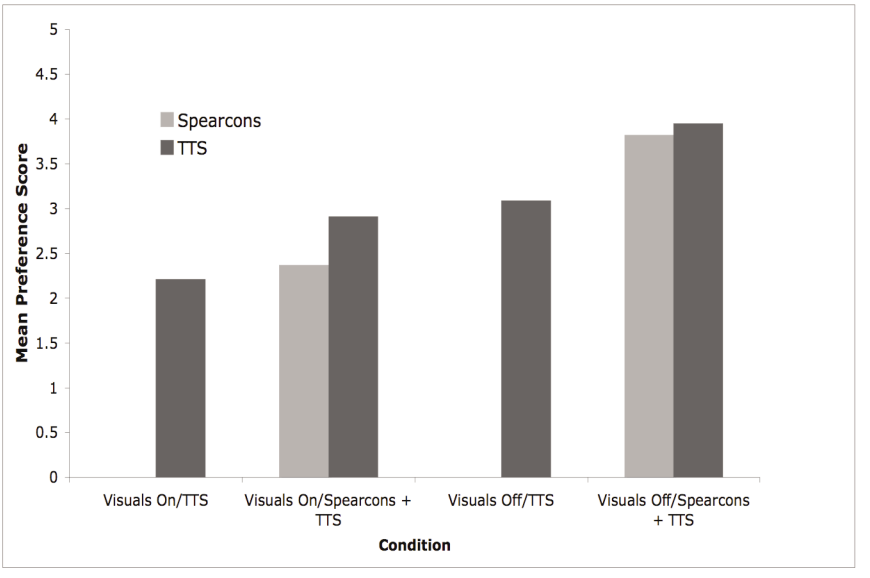

Fig. 4. Mean aggregate subjective preference scores, 5 being the highest possible score. TTS is given higher scores in the presence of spearcons.

4 Discussion

The performance results confirm many of the findings in the study by Palladino & Walker [1], allowing us to generalize the utility of spearcons as part of auditory menus from the desktop to the mobile phone. Conditions with visual cues led to faster responses, as compared to conditions with only auditory cues. This is understandable, given that the visual list allows for fast look-ahead. With the visuals on, the type of audio cues did not matter. That is, adding spearcons did not negatively impact performance, even though the spearcons add approximately half a second to each audio cue. In fact, even the silent (visuals only) condition was no different from the TTS and spearcons conditions, when the visual list was presented. It is likely the case that with the visuals on participants are moving through the list about as fast as possible by relying largely on the visual interface. Practice does not have much of an impact, supporting the interpretation that this is a highly practiced task. Adding the audio at least does not slow down performance when the visuals are on.

When the visuals are off, overall performance was slower than when visuals were on (see the top lines in Figure 2). However, with a little practice, performance in the audio-only conditions improved, and closed in on the conditions with visuals on (see the narrowing of the gap between the top lines and the bottom three lines, in Figure 2, from Block 1 across to Block 10). This bodes well for the use of auditory menus, even for users with little or no experience with audio-only interfaces.

Within the pair of audio-only conditions, it is interesting to note that TTS-alone initially led to faster performance than spearcons+TTS, but this difference went away by Block 10. In Block 1, it is likely the case that because the spearcons were prepended to the TTS for each item, participants took the time to listen to both cues

before making a selection, rather than focusing strictly on the spearcon to take advantage of its cuing capability. From the open-ended comments from participants, it appeared that they would hold down the arrow key to scroll quickly to the necessary item, then listen to the entire auditory cue and make the selection as needed. This showed that they made very little use of the auditory cue and relied mainly on their recollection of the alphabetical list organization. This would explain why a previous study by Palladino & Walker [10] showed a significant difference in the spearcons and TTS conditions while testing shallow *two-dimensional* menus. In that study, participants needed to listen to each menu item before proceeding to the next, since they could not predict what was coming. It was not beneficial for them to hold down the arrow key each time as they did in the present study with a deeper menu structure, as that would lead them to miss the necessary cue. However, as they became more familiar with both the list and the audio cues, participants here relied on the spearcons more. We know this because the overall performance times were comparable in the spearcons+TTS and TT-only conditions. That is, if they listened to, say, 1000 ms of audio for each menu item, then in the spearcons case this means about 250 ms of spearcon, 250 ms of silence, and 500 ms of TTS. Without the spearcon, this means 1000 ms of TTS. Thus, with practice, listeners came to make item selection decisions without listening to very much of the TTS phrases. Indeed, spearcons contribute a lot to performance of the navigation task.

The preference questionnaire demonstrated the positive reception of auditory cues in the absence of visual cues, as both spearcons and TTS were rated positively in the no-visual condition. This shows that, in a setting where users must rely on sounds to complete a task, they are inclined to feel good toward the sounds given, regardless of format. However, when they can rely on the visual sense to guide them, they prefer not to hear any audio and may even be annoyed by the sound. Given that there were no performance differences in the three visuals on conditions (silent, TTS only, and spearcons+TTS) it is instructive to consider the subjective ratings as well as the performance measures. Taken together, then, it is clear that users must be provided with the option to turn off audio when visuals are available, and turn it on only when it is perceived as desired and/or necessary. One additional caveat is that the audio quality needs to be optimized. Several participants commented on having trouble deciphering the audio cues for both spearcons and TTS. It is be important not to discount the interaction modality as a whole, simply due to a less-than-optimal implementation. While we are confident that the sounds here were generally acceptable and intelligible, the TTS could certainly be produced with higher quality algorithms. This would also improve the quality of the spearcons, since they are derived from the TTS sound files.

The general receptiveness of listeners to audio cues to aid navigation in a no-visual context supports further research into auditory menu design and deployment. In particular, it would be interesting to test how spearcons are perceived in a two-dimensional menu study, where they have shown improved performance over TTS alone. That is, what happens when both the preference and performance cues support spearcon use?

Most interestingly, although preference ratings for TTS were consistently higher than spearcons, the TTS ratings were even higher in the presence of spearcons. That is, adding spearcons to TTS seemed to enhance the ratings of the TTS. It is possible that listeners considered the spearcons+TTS menus to be more sophisticated or perhaps interesting, and this was rated as preferable. This has great implications for

designing with spearcons. While not harming overall user performance, spearcons can provide another layer to the user experience of audio menu navigation, one that encourages positive receptiveness to a new system.

5 Future Work

Future studies are focusing on the use of spearcons in audio-dependent contexts, where the participants cannot devote their full attention to the visual cue. In particular, we will be looking at task performance while a participant is simultaneously working on a visually and cognitively distracting task. This will be tested both in a desk setting and in a mobile one, where the user is walking on a designated route. We will be looking for effects on performance as well as subjective preference feedback from those involved. And, of course, we are extending these studies to participants with vision impairments, as they will be the primary users of (non-visual) advanced auditory menus, enhanced with whatever cues make the interfaces more effective and more pleasing to use.

References

1. Palladino, D., Walker, B.N.: Efficiency of spearcon-enhanced navigation of one-dimensional electronic menus. In: Proceedings of the International Conference on Auditory Display (ICAD 2008), Paris, France (2008)
2. Nees, M.A., Walker, B.N.: Auditory Interfaces and Sonification. In: Stephanidis, C. (ed.) The Universal Access Handbook, pp. TBD. Lawrence Erlbaum Associates, New York (in press)
3. Gaver, W.W.: Auditory Icons: Using Sound in Computer Interfaces. In: Human-Computer Interaction, vol. 2, pp. 167–177 (1986)
4. Blattner, M.M., Sumikawa, D.A., Greenberg, R.M.: Earcons and icons: Their Structure and Common Design Principles. In: Human-Computer Interaction, vol. 4, pp. 11–44 (1989)
5. Walker, B.N., Nance, A., Lindsay, J.: Spearcons: Speech-based Earcons Improve Navigation Performance in Auditory Menus. In: Proceedings of the International Conference on Auditory Display (ICAD 2006), London, England, pp. 63–68 (2006)
6. Palladino, D., Walker, B.N.: Learning rates for auditory menus enhanced with spearcons versus earcons. In: Proceedings of the International Conference on Auditory Display (ICAD 2007), Montreal, Canada, pp. 274–279 (2007)
7. Asakawa, C., Takagi, H., Ino, S., Ifukube, T.: Maximum Listening Speeds for the Blind. In: Proceedings of the International Conference on Auditory Display (ICAD 2003), Boston, MA (2003)
8. Leplatre, G., Brewster, S.: Designing Non-Speech Sounds to Support Navigation in Mobile Phone Menus. In: Proceedings of the International Conference for Auditory Display (ICAD 2000), Atlanta, GA, pp. 190–199 (2000)
9. Hereford, J., Winn, W.: Non-Speech Sound in Human-Computer Interaction: A Review and Design Guidelines. Journal of Educational Computing Research 11, 211–233 (1994)
10. Palladino, D., Walker, B.N.: Navigation efficiency of two dimensional auditory menus using spearcon enhancements. In: Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society (HFES 2008), New York, NY, September 22-26 (2008)