"Spindex": Accelerated Initial Speech Sounds Improve Navigation Performance in Auditory Menus

Myounghoon Jeon and Bruce N. Walker
Sonification Lab, Georgia Institute of Technology
Atlanta, Georgia, USA 30332-0170
mh.jeon@gatech.edu, bruce.walker@psych.gatech.edu

Users interact with mobile devices through menus, which can include many items. Auditory menus can supplement or even replace visual menus. Unfortunately, little research has been devoted to enhancing the usability of large auditory menus. We evaluated a novel auditory menu enhancement called a "spindex" (i.e., speech index), in which brief audio cues inform the user where she is in a long menu. In the current implementation, each item in a menu is preceded by a sound based on the item's initial letter. 25 undergraduates navigated through an alphabetized contact list of 50 or 150 names. The menu was presented with text-to-speech (TTS) alone, or TTS plus spindex, and with the visual menu displayed or not. Search time was faster with the spindex-enhanced menu, especially for long lists. Subjective ratings also favored the spindex. Results are discussed in terms of theory and practical applications.

INTRODUCTION

Users interact with many electronic devices through a menu system, and the menus themselves can include a great number of items (e.g., up to 30,000 songs on some portable music players[1]). Navigating these long menus requires special design considerations. Given the small (or non-existent) screen real estate available for a visual menu, and the challenges associate with using a visual menu in a mobile context (i.e., while walking, cycling, driving, or with the device in a pocket), an auditory menu is often an appropriate (or the only) way to present the menu items. Auditory menus are also very useful for users who are unable to use a visual menu due to temporary or permanent vision impairment (see Yalla & Walker, 2008 for an overview of auditory menus).

When an auditory menu is used, instead of—or in addition to—a visual menu, the same fast and accurate navigation requirements apply as for any menu interface. From this perspective, researchers have tried to improve the mobile systems by adding auditory cues. Device categories range from mobile phones (Brewster, Leplatre, & Crease, 1998; Gaver, 1986; Leplatre & Brewster, 2000; Palladino & Walker, 2007, 2008a, 2008b, Vargas & Anderson, 2003; Walker, Nance, & Lindsay, 2006) and PDAs (Brewster & Cryer, 1999; Klante, 2004), to wearable computers (Brewster, Lumsden, Bell, Hall, & Tasker, 2003; Wilson, Walker, Lindsay, Cambias, & Dellaert, 2007). While much of the work has focused on using or improving the speech parts of an auditory menu (typically generated by text-to-speech, or TTS; see Yalla and Walker, 2008), there have also been examples of non-speech sound cues successfully being used to help menu navigation (Brewster 1998; Palladino & Walker, 2007, 2008a, 2008b; Walker, Nance, & Lindsay, 2006). The present report reviews some of the non-speech sound cues previously used in auditory menu navigation and presents evidence on improvements provided by including a novel menu enhancements called a *spindex*, which is an auditory index based on speech sounds.

[1] Apple iPod classic, 120GB, holds up to 30,000 songs or 25,000 photos, as cited at http://www.apple.com/ipodclassic/

Types of Non-speech Auditory Cues

There have been three main approaches to enhancing the basic TTS used in most auditory menus. These all tend to include adding sound cues before or concurrent with the spoken menu items. The main types of enhancement cues are auditory icons, earcons, and spearcons (see Walker et al., 2006 for a longer review). In addition to these, we introduce the concept of the spindex.

Auditory Icons

Auditory icons (Gaver, 1986) are audio representations of objects, functions, and events. They are caricatures of naturally occurring sounds such as bumps, scrapes, or even files hitting mailboxes. As caricatures, auditory icons capture an object's essential features, by presenting a representative sound of the object. Auditory icons can represent various objects in devices more clearly than other auditory cues because the relation between a source of sound and a source of data is more natural than others. For example, a typing sound can represent a typewriter. Thus, auditory icons typically require little training and are easily learned. Auditory icons are also suited for presenting dimensional data such as the magnitude of some value. Moreover, they can categorize objects into distinct families. On the other hand, it is sometimes difficult to match all functions of devices with proper auditory icons. For example, it may be difficult to create a sound that clearly conveys the idea of "save" or "search" or "unit change" (Palladino & Walker, 2007, 2008a). Thus, auditory icons are of limited use in practical electronic devices, and are discussed here only for historical context.

Earcons

Earcons (Blattner, Sumikawa, & Greenberg, 1989) are non-verbal audio representations which are implemented in the user interface to provide information to the user about some objects, operations or interactions. Earcons are typically

composed of musical motives, which are short, rhythmic sequences of pitches with variable intensity, timbre and register. Since earcons use an arbitrary mapping between sound and object, they can be analogous to a language or a symbol sign. This arbitrary mapping between earcon and represented item means that earcons can be applied to any type of menu; that is, earcons can represent basically any concept. On the other hand, this flexibility provides a weakness because the arbitrary mapping of earcons to concepts requires training. They can also represent hierarchical menus by logically varying musical attributes. However, when a new item has to be inserted in a fixed hierarchy (e.g., adding a new name to a contact list), it might be difficult to create a new branch sound.

Spearcons

Spearcons are brief sounds that are produced by speeding up spoken phrases, even to the point where the resulting sound is no longer comprehensible as a particular spoken word (Walker et al., 2006). These unique sounds are analogous to fingerprints because of the acoustic relation between the spearcons and the original speech phrases.

Spearcons are easily created by converting the text of a menu item to speech via text-to-speech. This allows the system to cope with dynamically changing items in menu items. For example, the spearcon for "Save" can be extended into the spearcon for "Save As." Or, if a new name is added to a contact list, the spearcons can be created as needed, even on the fly. Also, spearcons are easy to learn whether they are comprehensible as a particular word or not, because they derive from the original speech (Palladino & Walker, 2007).

Spindex: An Auditory Index Based on Speech Sounds

A spindex is created by associating an auditory cue with each menu item, in which the cue is based on the pronunciation of the first letter of each menu item. For instance, the spindex cue for "Apple" would be a sound based on the spoken sound "A". Note that the cue could simply be the sound for "A", but could also be a derivative, such as a speeded-up version of that sound, reminiscent of the way spearcons are based on—but not identical to—speeded-up speech sounds. The set of spindex cues in an alphabetical auditory menu is analogous to the visual index tabs that are often used to facilitate flipping to the right section of a thick reference book such as a dictionary or a telephone book.

When people search (visually or auditorily) a long list such as in a mobile phone address book or MP3 player, the process can be divided into two stages. One is rough navigation and the other one is fine navigation (Klante, 2004). In rough navigation, users pass or jump the non-target alphabet groups by glancing at the initials. For example, users quickly jump to the "T" section. Then, once users reach a target zone and begin fine navigation, they check where they are and cautiously fine-tune their search. Klante (2004) found that users' navigation strategies could be identified as rough navigation and fine navigation in 3-D audio environment as well. With auditory cues, people cannot jump as easily, given the temporal characteristics of spoken menu items. However, they still want to pass the non-target alphabetical groups as fast as possible. If a sound cue is sufficiently informative, users do not need to listen to the whole TTS phrase (Palladino & Walker, 2007). The initials of the alphabet of the list can give enough information to users when sorting out the non-target items. The benefit of a cue structure like a spindex is most likely realized more clearly in long menus with many items in many categories or sections. Given that they are likely most useful in long menus, it is fortuitous that spindex cues can be generated quickly (on the fly) by TTS engines, and do not require the creation and storage of many additional audio files for the interface. This is an important issue for mobile devices which, despite increasing storage for content file, are not designed to support thousands of extra files for their menu interface. Finally, because spindex cues are part of the original word and thus are natural sounds (based on speech), they are expected to require little or no training.

Menu Navigation Performance Improvement by Non-Speech Auditory Cues

Performance improvement by the addition of auditory cues in menu navigation tasks has been studied by several metrics such as the measurement of reaction time, the number of key presses, accuracy, and error rate. In earlier work, earcons have shown superior performance compared to non-organized sound (Brewster, Wright, & Edwards, 1992) or no-sound in a desktop computer (Brewster, 1997), in a PDA (Brewster & Cryer, 1999) and in a mobile phone (Leplatre & Brewster, 2000). Further, musical timbres were more effective than simple tones (Brewster et al., 1992). Also, in a hierarchical menu experiment, participants with earcons could identify their location with over 80% accuracy. These findings showed that the logic of earcons is promising to apply them for hierarchy information (Brewster, Raty, & Kortekangas, 1996). Recent research on the addition of auditory scroll bars has demonstrated the potential benefits of applying earcons proportionally to each group of list items. The results showed reducing error rate in target search (Yalla & Walker, 2008).

Spearcons have recently shown promising results in menu navigation tasks. Walker et al. (2006) demonstrated that adding spearcons to a TTS menu leads to faster and more accurate navigation than TTS only, auditory icons + TTS, and earcons + TTS groups. Spearcons also improved navigational efficiency more than menus using only TTS or no sound when combined with visual cues (Palladino & Walker, 2008a, 2008b). According to Palladino and Walker (2008a), in their visuals-off condition, the mean time-to-target with spearcons + TTS is shorter than that with TTS only, despite the fact that adding spearcons makes the total system feedback longer.

Subjective Evaluation on Auditory Cues

Long or loud sounds may easily be annoying and can disturb other people's work. Therefore, using sounds in applications and devices should be carefully considered, especially with respect to duration, amplitude, and aesthetic quality. However, at present, there are only a few studies examining these issues for auditory menus. There are some studies on polarity preference of sound (Walker, 2002; Yalla & Walker, 2008), and some studies have investigated preference or annoyance of earcons (Leplatre & Brewster,

2000; Helle, Leptare, Marila, & Laine, 2001; Marila, 2002). Earcons seem to have more aesthetic aspects because they have musical motives. Nevertheless, frequently played sounds in devices can make users annoyed. A subsequent study (Helle, Leptare, Marila, & Laine, 2001) investigated the subjective reactions of users who used sonified mobile phones. Users did not prefer the sonified mobile phone because it was disturbing and annoying. One possibility is that the sounds of the study were too complex. Marila (2002) demonstrated that a simpler sound was preferred and enhanced performance more than a complex one. The alternative explanation about low preference of sound is that the sound quality was relatively low. Helle et al. (2001) disclosed their limitation of using only the beep sound. Nowadays, this quality limitation has been overcome as technology develops. However, there are many questions remaining to be solved on aesthetics, preference, and annoyance. Recent work has begun to study the subjective improvements to auditory menus from spearcons and other similar enhancements (Walker & Kogan, 2009). Now, in a similar vein, spindexes also have the potential to gain users' acceptance because they are natural, simple and short.

In the present study, then, undergraduate participants navigated auditory menus with TTS-only and TTS + spindex menus to examine whether adding a spindex would improve navigation efficiency and preference. We predicted that target search time for TTS + spindex would be shorter than that of TTS alone. We also predicted that spindex-enhanced menus would score higher than plain TTS on subjective ratings.

METHOD

Participants

25 undergraduate students (14 female; mean age = 20.4 years) participated in this study for partial credit in psychology courses. They reported normal or corrected-to-normal vision and hearing, signed informed consent forms, and provided demographic details about age and gender.

Apparatus

Stimuli were presented using a Dell Optiplex GX620 computer, running Windows XP on a Pentium 4, 3.2 GHz processor and 1 GB of RAM. An external Creative Labs Soundblaster Extigy sound card was used for sound rendering. Participants listened to auditory stimuli using Sehnheiser HD 202 headphones, adjusted for fit and comfort. A 17" monitor was placed on a table 40cm in front of the seated participant.

Stimuli

Two phonebook lists were composed. The short list had 50 names and the long version contained 150 names. These names (e.g., "Allegra Seidner") were created using a random name generator (http://www.xtra-rant.com/gennames/ & http://www.seventhsanctum.com). Visual stimuli consisted

Figure 1. Screen capture of mobile phone simulation with name list used to collect data.

of a mobile phone simulation with this list of names displayed in alphabetical order by first name in a "window (W*H=5cm*6 cm)" ten at a time. The particpant was able to scroll downward and upward in the simulated phone menu by pressing arrow keys on the keyboard (see Figure 1). The enter key was used to select the appropriate menu item. For both the auditory and visual components, if the participant reached the top or bottom of the list, the list did not wrap around. The experiment was built using Macromedia Director MX.

Text- to- Speech

TTS files (.wav) were generated for all of the names using the AT&T Labs Text-To-Speech Demo program with the male voice 'Mike-US English' (http://www.research.att.com/~ttsweb/tts/demo.php).

Spindex Cues

Spindex cues were also created by the AT&T Labs TTS Demo program. Each spindex cue consisted of only one syllable, pronouncing one of 26 letters of the alphabet. Spindex cues used in the address book were presented before the TTS cues, with 250ms interval between them (similar to spearcons, Palladino & Walker, 2008a; 2008b). If a participant pressed and held the up or down arrow key, the spindex cues were accelerated by the program. That is, the initials of the names were played without interval, in a preemptive manner.

Design and Procedure

A split plot design was used in this experiment. There were three within-subjects variables and one between-subjects variable. The within-subjects variables included list length (Short and Long), block (1~4) and auditory cue type (TTS only and TTS + spindex). The between-subjects variable was visual type (On and Off). The overall goal of the participants was to reach the target name in the address book menu as fast as possible, and press the enter key. There were no practice trials before the experiment blocks. The experiment was composed of four blocks in each condition. One block included 15 trials of different names as targets. All participants experienced the same procedure for each block, regardless of the assigned menu display conditions.

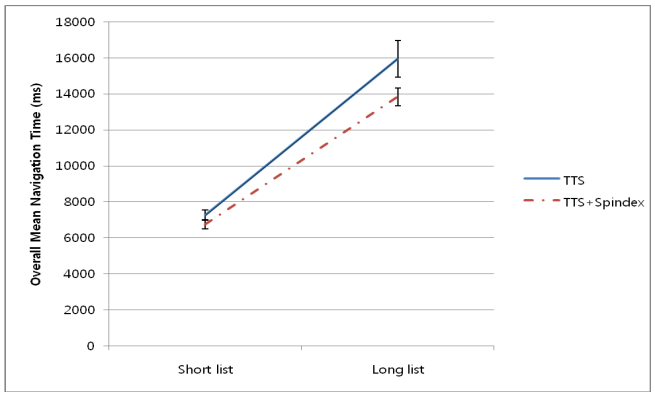

Figure 2. Interaction Effect of List Length & Auditory Cue type

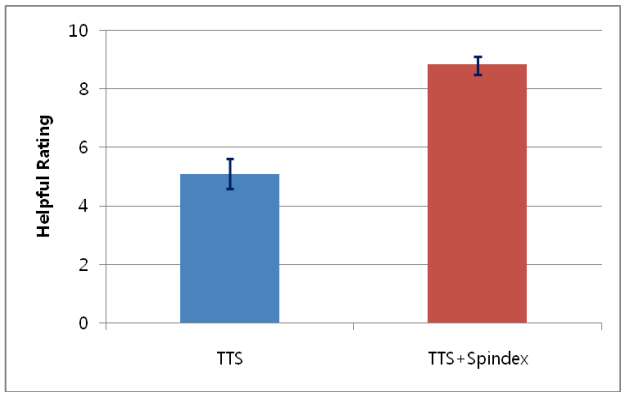

Figure 3. Helpfulness Rating Results

The order of appearance of the condition was counterbalanced within and between participants. A simulated mobile phone address book menu was presented that contained items constructed with auditory and visual representations.

On each trial, the target name was presented visually on the top of the computer screen. In the visuals-off group, the address book list was not shown, but the target name was still presented visually. When the participants first pressed the up or down arrow key, the timer started. Participants navigated through the menu system to find the assigned target name. Pressing the enter key worked to indicate selection of the requested target and recorded the end time. This procedure was repeated for all 15 names in the block. Participants were then shown a screen that indicated that the next block of 15 trials was ready to start. After four conditions, participants filled out a short questionnaire. An eleven-point Likert-type scale was used for the self-rated levels of *appropriateness*, *helpfulness*, *likability*, *fun*, and *annoying* with regard to auditory cues.

RESULTS

The results are depicted in Figures 2 and 3. In particular, Figure 2 shows the interaction between list length and auditory cue type. These results were analyzed with a 2 (Visual types) x 2 (Auditory cue types) x 4 (Blocks) x 2 (List types) repeated measures analysis of variance (ANOVA). The analysis revealed that participants searched significantly faster in the visuals-on condition (M = 8473.623, SD = 640.39) than in the visuals-off condition (M = 13423.08, SD = 666.54), $F(1, 23)$ = 28.67, $p < .001$, $\eta_p^2 = .56$. Participants in the TTS + spindex condition (M = 10292.28, SD = 325.83) searched faster than those in the TTS only condition (M = 11606.42, SD = 637.95), $F(1, 23) = 10.04$, $p < .05$, $\eta_p^2 = .30$. Also, main effect for block was statistically significant, $F(3, 69) = 9.25$, $p < .001$, $\eta_p^2 = .29$. In addition, a short list (M = 7003.12, SD = 238.31) led to significantly shorter search times than a long list (M = 14895.58, SD = 730.88), $F(1, 23) = 190.15$, $p < .001$, $\eta_p^2 = .89$. The interaction between list length and auditory cue type was also significant, $F(1, 23) = 8.95$, $p < .05$, $\eta_p^2 = .28$ (see Figure 2). This interaction reflects the fact that the spindex was more helpful in the long list than in the short list.

In the subjective data (Figure 3), the spindex-enhanced menu had a higher helpfulness score (M = 8.84, SD = 1.25) than the TTS only menu (M = 5.08, SD = 2.72). Paired-samples t-test indicated this difference was statistically reliable, $t(24) = -5.914$, $p < .001$. For the other scores (appropriateness, likability, fun, and annoying), there was no significant difference between the two groups.

DISCUSSION

Recently, some researchers have tried to improve auditory menu navigation through the implementation of extra auditory cues. The present study suggests that a new kind of auditory menu enhancement, which we call the spindex, can enhance navigation performance over TTS-only auditory menus. In this experiment, undergraduate participants showed better performance in the TTS + spindex condition than in the TTS only condition. The spindex enhancement effect was larger for longer auditory menus (150 items) than for relatively short menus (50 items). This is due to the fact that even small per-item enhancements lead to important and noticeable navigation times in long lists. Given that the more names in the list, the more helpful adding a spindex will be, this bodes very well for extemely long lists, such as those found in MP3 players.

Beck and Elkerton (1989) already showed that visual indexes could decrease search time with lists. The benefit of an auditory index (spindex) can be explained by the users' different strategies in the search processes: In the rough navigation stage, users exclude non-targets until they approach the alphabetical area including the target. This is possible because they already know the framework of alphabetic ordering and letters. Thus, during this process, they do not need the full information about the non-targets. It is enough for them to obtain only a little information in order to decide whether they are in the target zone or not. After users perceive that they reach the target zone, they then do need the detailed information to compare it with the target. Between these processes, the spindex-enhanced auditory menu can contribute significant per-item speedups in the rough search stage, then still support detailed item information via the TTS phrase, in the final search stage.

The fact that participants gave higher scores to the spindex menu on the subjective rating (helpfulness scale) indicated that they also felt that the spindex was helpful in the navigation. Of course, it is reassuring to display designers when objective performance (e.g., navigation time) and subjective assessments of appropriateness match! Even the few participants whose search times were not statistically

better in the spindex condition said that their strategy for navigation was to hear the initial alphabet sound of the names. This validates the spindex approach, even if in some cases it did not lead to a measurable improvement. At least, it did no harm. Of course, it may simply be the case that a reliable improvement from a spindex menu would come at even longer list lengths for these participants, or after further familiarity with the display; that remains to be seen. It is encouraging that using spindex cues requires little or no learning, which means a low threshold for new users. These advantages can increase the possibility of adoption in real devices.

Nevertheless, the details of adding a spindex need to be refined, in order to lessen annoyance—the annoyance ratings were relatively high (M = 6.24 out of 10, SD = 2.39). Thus, we are testing alternative designs such as attenuating the intensity of the spindex cues after the first one, and the use of a spindex cue only when crossing sub-list boundaries (e.g., for the first item starting with B, then the first C, and so on).

The use of mobile devices with smaller or non-existent screens, and longer and longer menus, is increasing. These devices are being used in more dynamic and mobile contexts, even with the device in a pocket. Enhancing auditory menus, with cues such as the spindex, can improve usability, accessibility, and subjective impressions of the devices and can ultimately provide essential information to users efficiently and pleasantly.

REFERENCES

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

Beck, D., & Elkerton, J. (1989). Development and evaluation of direct manipulation list. *SIGCHI Bulletin, 20(3),* 72-78.

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction, 4,* 11-44.

Brewster, S. A. (1997). Using non-speech sound to overcome information overload. *Displays, 17,* 179-189.

Brewster, S. A., Leplatre, G., & Crease, M. G. (1998). Using non-speech sounds in mobile computing devices. *Proceedings of the 1st Workshop on Human Computer Interaction with Mobile Devices*, Glasgow, UK.

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1992). A detailed investigation into the effectiveness of earcons. *Proceedings of the 1st International Conference on Auditory Display*, Santa Fe, USA, pp. 471-498.

Brewster, S. A. & Cryer, P. G. (1999). Maximising screen-space on mobile computing devices. *In summary proceedings of ACM CHI'99 (Pittsburgh, PA),* ACM Press, Addison-Wesley, pp 224-225.

Brewster, S., Lumsden, J., Bell, M., Hall, M., & Tasker, S. (2003). Multimodal 'eyes-free' interaction techniques for wearable devices. *CHI 2003*, Florida, USA, pp. 473-480.

Brewster, S., Raty, V. P., & Kortekangas, A. (1996). Earcons as a method of providing navigational cues in a menu hierarchy. *Proceedings of HCI'96*, pp. 167-183.

Gaver, W. W. (1986). Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction, 2,* 167-177.

Helle, S., Leplatre, G., Marila, J., & Laine, P (2001). Menu sonification in a mobile phone – a prototype study. *Proceedings of the 7th International Conference on Auditory Display*, Espoo, Finland.

Klante, P. (2004). Auditory interaction objects for mobile applications. *Proceedings of the 7th International Conference on Work With Computing Systems, WWCS2004*, Kuala Lumpur, Malaysia.

Leplatre, G. & Brewster, S. A. (2000). Designing non-speech sounds to support navigation in mobile phone menus. *Proceedings of the 6th International Conference on Auditory Display*, Atlanta, GA, pp.190-199.

Marila, J. (2002). Experimental comparison of complex and simple sounds in menu and hierarchy sonification. *Proceedings of the 8th International Conference on Auditory Display*, Kyoto, Japan.

Palladino, D. K., & Walker, B. N. (2007). Learning Rates for Auditory Menus Enhanced with Spearcons versus Earcons. *Proceedings of the 13th International Conference on Auditory Display*, Montreal, pp. 274-279.

Palladino, D. K., & Walker, B. N. (2008). Efficiency of spearcon-enhanced navigation of one dimensional electronic menus. *Proceedings of the 14th International Conference on Auditory Display,* Paris, France.

Palladino, D. K., & Walker, B.N. (2008). Navigation efficiency of two dimensional auditory menus using spearcon enhancements, *Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society (HFES 2008)*, San Antonio, Texas.

Walker, B. N. (2002). Magnitude estimation of conceptual data dimensions for use in sonification. *Journal of experimental psychology: Applied, 8(4),* 211-221.

Walker, B. N., & Kogan, A. (2009). Spearcons enhance performance and preference for auditory menus on a mobile phone. *Proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction (UAHCI) at HCI International 2009*, San Diego, CA, USA, (19-24 July). pp. TBD.

Walker, B. N., Nance, A., & Lindsay, J. (2006). Spearcons: Speech-based earcons improve navigation performance in auditory menus. *Proceedings of the 12th International Conference on Auditory Display,* London, England.

Wilson, J., Walker, B. N., Lindsay, J., Cambias, C., & Dellaert, F. (2007). SWAN: System for Wearable Audio Navigation. *Proceedings of the 11th International Symposium on Wearable Computers (ISWC 2007)*, Boston, MA.

Yalla, P., & Walker, B. N. (2008). Advanced auditory menus: Design and evaluation of auditory scroll bars. *Proceedings of the 10th international ACM Conference on Assistive Technologies (ASSETS 08)*, Halifax, Nova Scotia, Canada.