earPod: Eyes-free Menu Selection using Touch Input and Reactive Audio Feedback

Shengdong Zhao[1], Pierre Dragicevic[1], Mark Chignell[2], Ravin Balakrishnan[1], Patrick Baudisch[3]

[1]Department of Computer Science
University of Toronto
sszhao, dragice, ravin@dgp.toronto.edu

[2]Department of Mechanical and Industrial Engineering
University of Toronto
chignell@mie.toronto.edu

[3]Microsoft Research
Redmond, WA 98052
baudisch@ microsoft.com

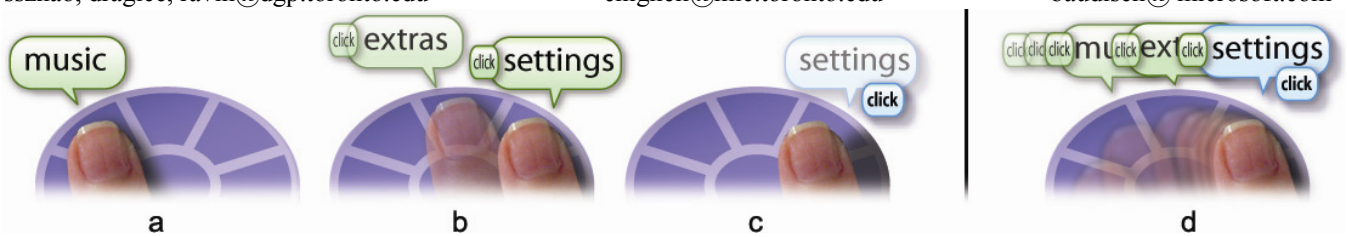

Figure 1. Using earPod. (a, b) Sliding the thumb on the circular touchpad allows discovery of menu items; (c) the desired item is selected by lifting the thumb; (d) faster finger motions cause partial playback of audio. Size of the touchpad has been exaggerated for illustration purposes.

ABSTRACT

We present the design and evaluation of earPod: an eyes-free menu technique using touch input and reactive auditory feedback. Studies comparing earPod with an iPod-like visual menu technique on reasonably-sized static menus indicate that they are comparable in accuracy. In terms of efficiency (speed), earPod is initially slower, but outperforms the visual technique within 30 minutes of practice. Our results indicate that earPod is potentially a reasonable eyes-free menu technique for general use, and is a particularly exciting technique for use in mobile device interfaces.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. Input devices and strategies;

Keywords: Gestural interaction, auditory menu

INTRODUCTION

The visual modality has long dominated HCI research. However, visual feedback is not always desirable or feasible due to a number of factors. 1) C*ompetition for visual attention*: in mobile scenarios (walking, running or driving), users need to pay attention to the environment and looking at the interface may be distracting or dangerous [9]. 2) *Absence of a visual display*: many devices, due to cost, space constraints or historical reasons, do not have a display (e.g.: telephone, iPod Shuffle). 3) *User disability*: Depending on the severity of the disability and the size and complexity of the display, people with visual disabilities may be unable to use a visual display. 4) *Inconvenience*: viewing the screen of a device will typically require taking the device out of the user's pocket or bag, which may be particularly undesirable in inclement weather. 5) *Reduction of battery life:* illumination of an LCD screen to allow visual navigation dramatically reduces battery life for electronic devices, also reducing the device's true mobility.

Designing for eyes-free use requires exploring other input and output modalities. Non-visual output modalities such as tactile feedback have been investigated [19]. However, the technology required to generate tactile feedback is typically complex. Auditory feedback would seem to be a more promising option, especially given existing technological support in mobile devices such as phones and media players.

Unlike graphical user interfaces (GUIs), where general interaction models (e.g., WIMP) have been extensively studied and used, there is not yet a similar "standard" for auditory interfaces. Due to the significant differences between the audio and visual channels, a straightforward translation from visual interfaces often fails at leveraging the full potential of the audio channel. Thus, designing general auditory interaction techniques that are intended to replace existing visual techniques requires careful thought and experimentation. In this paper, we focus on menu selection.

Designing an efficient hierarchical auditory menu system has long been recognized as a difficult problem [20, 32]. These challenges are revealed in the design of automatic interactive voice response (IVR) systems, which are often referred to as "touchtone hell" [43]. Incremental improvements to IVR menus have been proposed to ease user frustration [8, 20, 26, 31, 41], but despite these improvements, auditory menus are still harder to use than visual menus.

It has been argued that the difficulty of navigating auditory menus is fundamentally rooted in the serial and temporal nature of auditory information, and using the auditory channel alone is unlikely to ease users' frustration. In order to overcome these difficulties, Yin and Zhai [43] have proposed an augmented IVR system with synchronized visual feedback. However we argue that audio-only menus, at least for static menus of reasonable size, can be efficiently used provided that they are redesigned appropriately.

CHI 2007, April 28-May 3, 2007, San Jose, California, USA.

We present a touch-based auditory menu technique called earPod (Figure 1 & Figure 2). earPod provides users with audio feedback that is synchronously linked to touch input. earPod is intended to allow users to discover menus at their own pace. Seamless transition from novice to expert use is facilitated with an absolute radial menu layout that allows direct access to known items. Spatialized audio feedback additionally reinforces the mappings of items to a circular touch-sensitive surface. Our earPod interface requires only standard inexpensive hardware, i.e., a circular touchpad and stereo headphones (Figure 2). This would make it easy to deploy earPod with existing touchpad-equipped devices such as iPods, PDAs, and laptops, and with cell phones and desktop computers if retrofitted with an external touchpad.

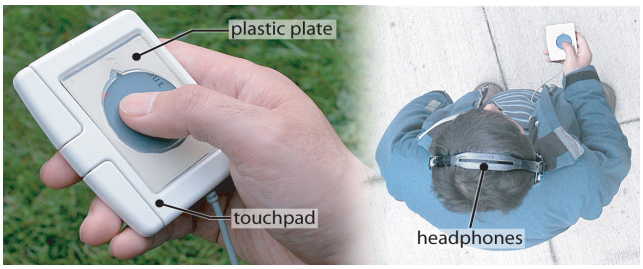

Figure 2. Our earPod prototype uses a headset and a modified touchpad

In this paper, we review related work and provide a detailed description of the earPod technique and its design rationale. We then present an experiment comparing earPod with an iPod-like visual menu. We conclude with a discussion of our results and directions for future work.

RELATED WORK

Numerous researchers have inspired our work. There is a large body of work on audio-based icons (e.g.: Gaver's auditory icon [14], Blattner et al. [6] and Brewster et al.'s [10] earcons), auditory interfaces (e.g.: Schmandt and colleagues [3, 22, 34-37, 39, 40], Mynatt and colleagues [23]) and visual menu design (summarized in Kent Norman's book [24]). Relatively fewer studies have focused on the interaction design of auditory menus. Besides the previously mentioned research on IVR systems [20, 26, 31, 41, 43], Brewster [8] also tried using non-speech audio to provide navigational cues on voice menus. Recently, the emergence of mobile computing has inspired researchers to rethink the interaction model of audio command selection. Pirhonen et al. [29] investigated the use of simple gestures and audio-only feedback to control music playback in mobile devices. Brewster et al. [7] also investigated the use of head gestures to operate auditory menus. Both techniques have demonstrated effectiveness in the mobile environment. However, they have only been investigated with a very limited number of commands. For example, the head gesture menu Brewster et al. created used only four options, which is insufficient for the wide range of functionality that exists in today's devices.

Recently, Liao et al. [18] presented a multimodal pen prototype to allow users to perform menu selection without digital displays. earPod shares a number of key features with their technique, including a circular menu layout and interruptible speech. However, Liao et al. focused on paper-based pen interfaces and their findings are not easily generalizable to scenarios of pure eyes-free and mobile interaction. Mobile environments are often unstable and need to overcome issues such as limited hand availability [27], making it difficult to rely on pen operations. Furthermore, in Liao et al.'s approach, auditory feedback is presented and evaluated as part of a multimodal technique that also includes visual (ink and LEDs) and tactile feedback. Menus using only auditory feedback deserve further investigation. In particular, it will be valuable to determine if audio feedback alone is efficient and effective when combined with touch input, and the extent to which this combination can be exploited in a range of eyes-free usage scenarios. These considerations motivate and differentiate our present work.

EARPOD WALKTHROUGH

The earPod technique is designed for an auditory device controlled by a circular touchpad whose output is experienced via a headset (Figure 2), as is found, for example, on an Apple iPod. Figure 3 shows how the touchpad area is functionally divided into an inner disc and an outer track called the *dial*. The dial is divided evenly into sectors, similar to a pie [11] or Marking Menu [17, 44, 45].

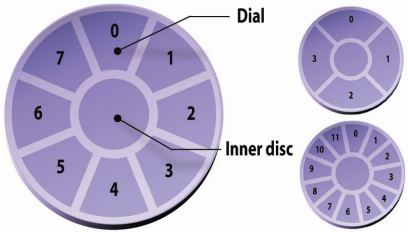

Figure 3. The functional areas of earPod's touchpad. Up to 12 menu items can be mapped to the track. The inner disc is used for canceling a selection.

Our technique is illustrated in Figure 1. When a user touches the dial, the audio menu responds by saying the name of the menu item located under the finger (Figure 1a). Users may continue to press their finger on the touch surface, or initiate an exploratory gesture on the dial (Figure 1 b). Whenever the finger enters a new sector on the dial, playback of the previous menu item is aborted. Boundary crossing is reinforced by a click sound, after which the new menu item is played. Once a desired menu item has been reached, users select it by lifting the operating finger, which is confirmed by a "camera-shutter" sound (Figure 1c). Users can abort item selections by moving their finger to the center of the touchpad and releasing it. If a selected item has submenus, users repeat the above process to drill down the hierarchy, until they reach a desired leaf item. Users can skip items rapidly using fast dialing gestures (Figure 1d).

earPod is designed to allow fast expert usage. As users gain knowledge of the menu configuration through practice, they tend to use brief corrective gestures (Figure 1b) instead of large exploratory ones (Figure 1d). Eventually, as users remember the exact locations of desired menu items, they select these items by directly tapping on them.

DESIGN RATIONALE

We now discuss the design rationale behind earPod. We first analyze the weaknesses of traditional voice menu designs. In particular, we argue that IVR systems are inconvenient not only due to the serial and temporal nature of sound, but also because of the sequential dialog between the system and users. These two factors combine to cause an unbalanced interaction model between input and output (Figure 4, left), which we discuss in detail below.

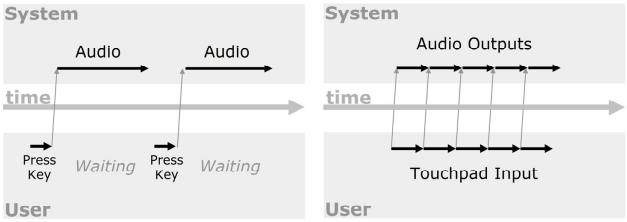

Figure 4. (left) The asynchronous interactive model of standard voice menus (IVR). (right) The synchronous interactive model of earPod.

Lessons from Previous Auditory Systems

The limitations of the audio modality compared to the visual one are well understood. The visual modality allows displaying a list of choices instantaneously and persistently. As a result, users can easily scan and compare visual menu items back and forth at their own pace, without having to commit any to memory [43]. In contrast, using audio to convey a list of choices requires users to adapt to an imposed rate of presentation and rely on their short-term memory. The presentation rate will often be either too slow (e.g., when information is irrelevant) or too fast (e.g., when information is critical) for the user.

Traditional IVR systems typically suggest a list of options, then prompt the user for an answer. They follow a *sequential* or *asynchronous dialog* model (Figure 4, left). A sequential dialog model is useful when the user needs assistance for a complex task. However, it is not appropriate for simple and repetitive tasks such as menu selection, and is particularly irritating when combined with speech output. The user often has to wait, with little or no control over the system. As Figure 4 (left) illustrates, after users press a key, they often wait several seconds to hear the entire audio feedback before initiating the next action. Such a wait is an important cause of the unpleasant user experience described as “touchtone hell” [43].

Many of today’s IVR systems allow quick access to specific menu items provided the user knows the exact code. However, because recovering from errors in sequential dialog systems is a costly process, using these features requires a great deal of self-confidence and practice. As a result, most intermediate users will listen to the whole list of options repetitively rather than taking the risk of hitting the wrong key. In fact, IVR systems are designed for both complete novices and complete experts but neglect the vast majority of users in between those extremes.

In his investigation of synchronous programming languages, Berry [5] contrasted the asynchronous interaction paradigm against the notion of a *reactive system*. In a reactive system, the user always has the initiative and never waits. Ideally, the system waits for user actions and reacts promptly whenever they occur. Such an approach has been widely adopted in modern direct manipulation interfaces. They give the user a positive *feeling of control* [38]. In addition, emphases on *proactive discovery* as well as *progressive acquisition of expert knowledge* are made [17, 38, 44, 45]. earPod demonstrates how auditory interfaces can benefit from similar design principles (Figure 4, right).

earPod’s Approach

To improve on the IVR paradigm, we made several design improvements.

Touch input: we chose to use a touchpad instead of a keypad for input. Touchpads arguably have a richer input vocabulary than keypads because they allow gliding gestures in addition to discrete taps. These gliding gestures allow browsing of menu items before confirming the selection. During exploration, the user’s finger is guided by the raised edges of the circular-shaped touchpad. Unlike a keypad with a fixed number of physical buttons, a touchpad allows flexible division of input area into arbitrary numbers of subsections (Figure 3, right), which can then be assigned to different commands. In addition, by using a circular touchpad, we remain compatible with existing devices such as the iPod, potentially allowing our technique to be installed simply as a software update.

Reactivity: instead of relying on an asynchronous communication model, we adopt a synchronous approach in which the system only reacts to the user’s actions (Figure 4, right). Menu items are mapped to physical areas on the touchpad and are played (as a voice cue) upon invocation, thus users have the ability to proactively discover available options at their own pace. When the information is uninteresting, they can skip to an adjacent area. Or they can listen to the entire message and repeat it if needed. By moving the finger back and forth, the user achieves the effect of "scanning and comparing" the menu items without having to commit them to memory. Such benefits were previously only provided by visual menus.

Interruptible audio: following the reactive interaction model, an auditory item is played back as soon as the finger reaches its location, regardless of whether the previous item is finished playing or not. We tested simultaneous playback (all items are played back entirely), interrupted playback (each new playback stops the previous one) and mixed approaches (previous playbacks fade out). Preliminary tests indicate simultaneous playback is confusing for users. To provide a much stronger feeling of reactivity, we eventually adopted interrupted playback.

It is especially useful to be able to promptly switch to a new item before the previous one finishes playing because users often understand partial audio messages[1]. Moreover, the amount of information needed to identify an option can be

[1] For example, if the user is looking for the item “apple”, “banana” can be rejected as soon as the syllable “ba” is heard.

further reduced as the user gains more information about the menu options. This allows the user to find an item faster than if all of the menu items had to be fully presented.

Non-speech audio: in addition to speech playback of menu items, we use non-speech audio to provide rapid navigational cues to the user. Non-speech audio has been shown to be effective in enhancing the graphical user interface [13, 14], and in improving navigation of non-visual information on mobile [9, 35] and IVR systems [8]. Short mechanical click sounds are played each time a boundary is crossed on the touchpad, in a way similar to the iPod's ClickWheel. These sounds are very helpful for separating the playback of new items from the playback of previous items. And even when the finger slides too fast for the speech audio to be heard, this mechanical sound gives a rough idea of the number of items or boundaries crossed. A "camera-shutter" sound is also used to confirm item selection.

Direct item access: even though the content of the entire menu can be played back sequentially, items can also be accessed directly. This is inspired by the design of pie or Marking Menus which lay out items radially. All items can thus be theoretically accessed in an equal amount of time. In our system, direct access can also be achieved by simply tapping the touchpad at the appropriate location whenever the user remembers an item's location. Such support for direct invocation is particularly beneficial in audio menus due to the slow rate of speech output. It allows users to completely bypass the audio playback, saving time.

Transition to expert use: in contrast to IVR systems, earPod provides a smooth transition from novice to expert in a way similar to Marking Menus. Each time users select an item using a dial gesture, they gain experience which should facilitate their transition to expert usage. In the beginning, the user tends to glide a longer distance to reach the desirable item (Figure 1d). As the user learns the absolute position of the target, the navigation path on the touchpad will be shortened (Figure 1b) and eventually approach direct invocation by tapping (Figure 1c). To distinguish between novice and expert usage, Marking Menus are typically implemented with a ~300 ms time-out before transitioning from marking to popup menu mode. However, the length of the timeout can change with different systems [18], and should also be adjusted to the needs of different users [44]. Instead of using a timeout, earPod eliminates the need for an artificially determined threshold by using interruptible audio and a self-discoverable touchpad layout. This transition is self discoverable, seamless, and arguably "natural".

Input / output mapping: we use spatialized audio to reinforce the user's cognitive mapping between menu items and spatial locations on the touchpad. For example, if the finger touches an item on the right side of the touchpad, the audio will be played back on the right side of the user (using binaural spatial cues). Consistency between the spatial knowledge obtained through finger exploration and the audio feedback is designed to help the user memorize the spatial layout of the items of interest.

IMPLEMENTATION DETAILS

To make earPod more accessible, our design only requires a circular touchpad as the input device and an ordinary stereo headset for output. Particular attention was paid to audio design as even minor details can have significant consequences on the usability of an auditory interface.

Input

The input device we used was a Cirque EasyCat USB external touchpad covered by a thin plastic overlay with a circular cutout, resulting in a circular touchpad. The radius of the touch-sensitive area is 19 mm and the overlay is 2 mm thick. The cutout includes a 3 mm notch on top, providing a tactile orientation cue to the user (Figure 2, right).

earPod was implemented in Java 1.5 and uses JNI to read raw absolute (x, y, pressure) touchpad data via a special driver supplied by Cirque.

Noise and inaccuracy were handled in two ways. First, involuntary movements occurring when the finger is lifted were filtered out. Second, a spatial hysteresis algorithm was applied to avoid inadvertently releasing the finger on an adjacent area. The algorithm works as follows: every time the finger crosses a boundary separating two functional areas of the touchpad, this boundary is slightly moved to the opposite direction. For example, if the finger moves to an area situated on the right, it will require a slightly larger movement to go back to the area on the left[2]. The same filtering methods were applied to the visual menu used in the experiment.

Output

All speech sounds used in earPod are human voices recorded in CD quality (16 bits, 44KhZ) using professional equipment. Since Java's native sound library adds a small but perceptible lag when playing sounds, we extended a real-time sound physics simulation library for use in earPod [12]. We also post-processed sound files after loading them to remove leading silences, and the sound signals were normalized to avoid the unpleasant effect of non-uniform playback volumes.

The third and final post-processing step was spatialization. In our current implementation, spatialization is simply achieved by manipulating mono voice streams to incorporate Interaural Time Differences (ITDs) and Interaural Intensity Differences (IIDs), the two major binaural cues for localizing sounds on the left-right axis [4]. Our ability to locate sounds on the two other axes is weaker and supporting it would require sophisticated signal filtering methods (such as Head Related Transfer Functions) as well as individual calibration [4]. Our technique does not preclude use of such methods, but simple left-right spatialization is better suited for mobile devices with limited computing power. Different spatialization methods can also be used in external speaker settings, such as home cinemas and cars.

[2] The accompanying video provides a visual demonstration of the hysteresis algorithm.

Interestingly, preliminary testing also showed that a simple left-right spatialization was enough to convey the illusion that items were laid out on a circle around the user's head. This suggests there may be a cross-modal integration effect between sounds and finger movement; although more research is needed to support this hypothesis (the Figure 2 may help in illustrating this point).

EXPERIMENT

Extensive research has been carried out on the configuration of radial menus [17, 44, 45]. In order to achieve acceptable speed and accuracy, it has been recommended that the circle used in such menus be divided up into 8 regions per menu level. We conducted pilot studies with breadth of 4, 8, and 12 regions, and with depths of 1 or 2 levels, and the results were consistent with the earlier findings; although 12 items are usable, 8 items or less per menu level tend to work best in terms of speed and accuracy.

Preliminary testing indicated that earPod menu selection was surprisingly easy and fast. As a result, we decided to compare the earPod approach with the commonly used visual selection approach implemented in the iPod (Figure 5, left). The iPod is a mature product that has been embraced by millions of consumers and has undergone several generations of iterative design. Similar to the earPod hardware, the iPod uses a circular touchpad for input, and allows navigation of alternatives by gliding the finger along the outer ring of the touchpad. As users glide the finger along the outer track of the ClickWheel (Figure 5, right), the corresponding items are highlighted. Users select the currently highlighted item by pressing the center button.

If the speed and accuracy of our input technique was comparable or only slightly worse than the iPod user interface, this would make our eyes-free technique a viable alternative to the visual interface of the iPod.

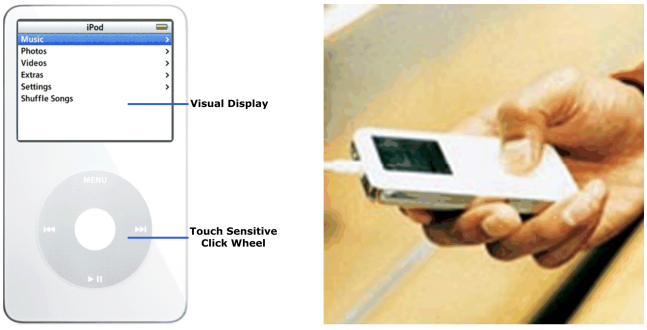

Figure 5. The iPod visual menu (left) and its interaction technique (right).

We were also interested in the transition from novice to expert performance, where experts were expected to tap directly on target items rather than move the finger around the touchpad's circumference.

To further investigate these issues, we conducted an experiment comparing earPod with an iPod-like visual menu. Since we were unable to receive the implementation details from Apple, the iPod linear menu was simulated using a circular touchpad connected to a notebook computer. Our implementation was identical to the recently released *iPod Nano* Media Player, apart from the fact that items were selected on finger release.

Participants

Twelve right-handed participants (3 female) ranging in age from 17 to 30 years (mean 24), recruited within the university community, volunteered for the experiment. None had previously used an iPod.

Apparatus

The experiment was conducted on a Dell Inspiron 8000 laptop running Microsoft Windows XP, with a 19" external LCD display. Input and output were handled as explained earlier in the implementation section.

Task and stimuli

To make the experiment more realistic, we chose to test our technique with a larger hierarchy than used in previous attempts. We opted for a 2-level hierarchy with 8 items per category, resulting in 64 items in total. We felt that this was a reasonable number of items to capture the common commands for many of today's audio-based devices. Preliminary studies, however, suggested that learning the location of 64 items would require more than one hour. To keep the duration of the experiment within reasonable limits, we decided to present only 16 of the 64 possible stimuli – 2 items per top-level menu. This still required participants to search the 8x8 item menu options and therefore made the selection process more demanding than a single 16-item menu. Limiting the number of possible stimuli, however, increased the frequency with which participants encountered each item – consistent with power law distributions observed for the frequency of use of command menus [42].

Following Miller [21], who used real world hierarchies as stimuli in his experiment, the categories we used were familiar words selected from hierarchies developed by KidsClick! [1] and Wikipedia [2]:

Clothing: Apron, Brief, Cloak, Coat, Dress, Hat, Jacket

Fish: Carp, Cod, Eel, Haddock, Pollock, Redfish, Salmon, Sardine

Instrument: Bassoon, Cello, Clarinet, Drums, Flute, Guitar, Organ, Piano

Job: Actor, Cook, Doctor, Driver, Farmer, Hunter, Lawyer, Soldier

Animal: Ants, Apes, Bats, Bears, Eagles, Zebras, Elephants, Horses

Color: Black, Blue, Grey, Green, Lime, Navy, Olive, Purple

Country: Brazil, China, Denmark, Egypt, England, Finland, France, Greece

Fruit: Apple, Banana, Cherry, Grape, Guava, Kiwi, Lemon, Mango

All words had one to three syllables and an audio duration of about 1 second. Although real menu items can contain several words, this setup captured the common case of single-word commands.

To assure that we would not favour the audio interface, we chose a visual presentation of stimuli for both conditions rather than an auditory presentation.

Procedure

Before their first trial, participants were instructed to put on the headphones and to hold the touchpad with their right hands as shown in Figure 2, leaving the thumb off the touchpad. Participants then pressed the spacebar using their left hand to start the trial. A visual stimulus (the item to select) was then displayed in the center of the screen as shown in Figure 6. Participants responded by bringing their thumb in contact with the touchpad and dragging the thumb in search of the target.

In the audio condition, participants heard the spoken names of each traversed menu item through their headphones. In the visual condition, menu items were displayed on the screen (see Figure 6). The selection process remained active as long as the participant's thumb remained in contact with the touchpad surface. Participants completed selections (either a menu or submenu item) by lifting the thumb off the touchpad. In both conditions, a short click sound was played whenever a selection was made. If a stimulus had two levels, participants went through the thumb-down → thumb-drag → thumb-up[3] process twice to finish the task. A trial was considered erroneous if any of the selected targets did not match the stimuli. In this case, participants were notified by a "mismatch" visual message . No additional feedback was given for successful trials. After each trial, a visual message in the center of the screen instructed participants to press the spacebar to proceed to the next trial.

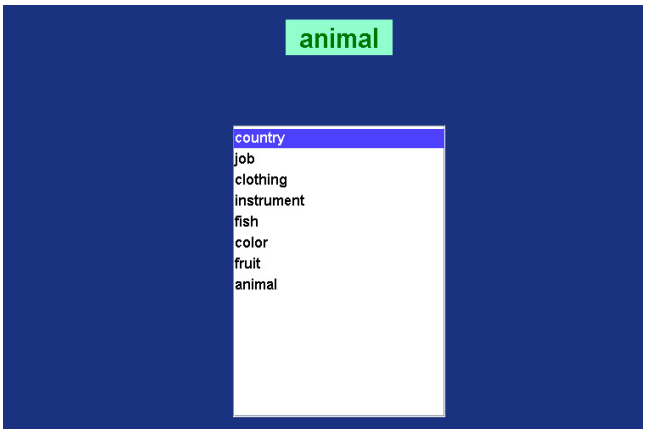

Figure 6. The lower half of the screen: the item to be selected from the menu is displayed at the center of the screen. The menu at the bottom is displayed only in the visual condition.

Design

A within-participants design was used. Participants were randomly assigned to two six-person groups. The first group performed the experiment with the earPod technique first, the second group with the iPod-like visual menu technique.

For each technique, participants made selections from 2 menu layouts: a single level menu containing the 8 categories, and a two-level menu with 8 items per level (8x8), containing all 64 items organized into the same 8 categories. We counter-balanced the technique, but used a single ordering of menu depth, from easy to difficult (first the 8-item menu, then the 8x8-item menu). The menu content and item orderings were chosen in advance and were identical for both techniques and for all participants:

Condition 8: all 8 possible stimuli were used.

Condition 8x8: as previously discussed in the "task and stimuli" section, to keep the experiment manageable, a subset of 16 of the possible 64 stimuli were chosen (two sub-items per top-level menu item, pre-selected before the experiment). Menu item presentation order was randomized across participants.

Participants were allowed to take breaks between trials. Breaks were enforced between different techniques and layouts. Before the experiment, participants received 5 minutes of training per interface using a different set of stimuli than the one used in the study. Each participant performed the entire experiment in one sitting which took approximately 90 minutes.

In summary, the design was as follows (excluding training):

12 participants x
2 techniques (audio and visual) x
(8+16) items for the 2 menu configurations (8 and 8x8) x
20 blocks
= 11520 menu selections in total.

Results

Repeated measures of analyses of variance were used to assess the effects of interface (audio vs. visual) on accuracy and selection time. For these analyses, the learning effect were assessed by grouping the 20 blocks of trials within each session into four groups of five contiguous blocks, which will be referred to as the four "time periods" below. Analyses were carried out across all four of the time periods to assess learning effects, and across the last time period (last five blocks) to assess the participants' performance with each technique after some training had taken place.

Accuracy

Overall, the audio technique had an accuracy of 92.1% in this study, while the visual technique yielded 93.9% accuracy. This difference was not statistically significant ($p>.05$). As might be expected, there was a significant effect for the number of menu levels ($F_{1,11}=21.16$, $p<.001$); accuracy with single level menus (94.2%) was higher than accuracy with two level menus (91.8%). None of the other main effects or interactions involving accuracy were significant when all four time periods were considered, nor when only the last time periods (last five blocks) were considered.

Selection Time

In each trial, selection time was measured as the duration from the appearance of the stimulus to the completion of the selection. As is typical with response time data, the distribution of raw times was positively skewed. During aggregation of the data (within participants and within each cell of the design), medians were used as measures of

[3] When users become experts in the audio condition, they skip the thumb-drag and only perform a tap to select an item.

central tendency in order to reduce the effect of potential outliers [33]. Means of these median values were then used to estimate average selection times across the participants.

There was no significant overall difference in speed between the audio and the visual technique ($F<1$, $p>.05$). As expected, however, it took significantly longer ($F_{1,10}=82.84$, $p<.001$) to complete selections in two level menus (averaging 3.4 seconds) than in one-level menus (1.9 seconds). There were also significant learning effects on response time across the experimental sessions. There was a main effect on overall learning ($p<.001$), and significant interactions between learning and number of menu levels ($p<.001$), and learning and technique ($p<.001$).

The learning effect with respect to the significant three-way interaction that was observed between time period, technique and number of menu levels ($F_{3,30}=7.75$, $p<.01$), is shown in Figure 7. The learning rate was faster with the audio menu than with the visual menu as illustrated in Figure 7 by the crossover in the curves that occurred for both menu levels. The audio menu was initially slower than the visual menu, but with experience, performance on the audio menu became faster. In the final time period (last five blocks of the twenty blocks) in the experimental session, the audio menu, at an average of 2.1 seconds, was significantly faster than the visual menu at an average of 2.5 seconds ($F_{1,10}=6.03$, $p<.05$).

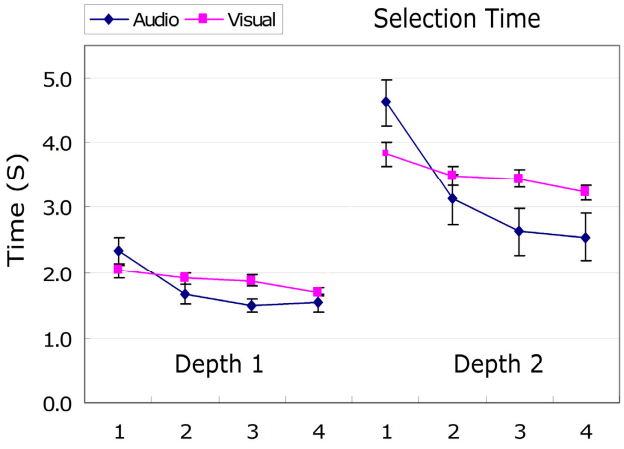

Figure 7. Selection times for the two techniques by number of menu levels and time period (1 time period = 5 contiguous blocks of trials).

Detailed Log Analysis

In addition to the accuracy and selection time data, we also maintained a detailed log of the exact movements made during the sessions. Using this log we were able to distinguish between glide selections, where participants traversed to the target region around the perimeter of the circular touchpad, and tap selections, where participants jumped to the target region and tapped it.

At the beginning of the experiment, all participants used the glide technique, with considerable variability in the details of the gliding traversals employed. The touchpad was divided into 8 zones as shown in Figure 3. From the beginning of the experiment, participants typically started by gliding to and around the target because they did not have the expertise to tap on the target immediately. For example, participant 1 went through the following sequence to find target item 5 on his first trial: zone 6 was touched first, and then he traversed (dragged) through zones 7, 1, 2, 3, 4, before finally settling on zone 5, and releasing his finger to select it. However, by the end of the session the participant was tapping on the target zones directly without any drags, indicating that expert performance had been reached. Similarly, for a 2-level selection, expert performance was indicated by using only 2 taps, avoiding the need to drag. However, not all the participants were able to reach expert performance in the experiment. Overall, 9 out of 12 participants reached expert performance for the single level menu, and 8 out of 12 participants reached expert performance in the two level menu. It should be noted that the only technique taught to the participants was gliding, thus 9 out of 12 participants independently discovered the tapping technique. Even for the 3 participants who did not reach expert behavior by the end of the session, examination of their detailed logs indicated that they used tapping in some of the experimental trials.

Subjective Preference

In terms of subjective preference, four participants expressed a preference for the audio menu, four preferred the visual menu (including the three participants who did not achieve expert performance), and four had no preference between the two techniques. This result seems promising, as the participants had no incentive to use an eyes-free technique under the given experimental settings (i.e., they were not required to walk or asked to perform a visually distracting task while making their menu selections). All participants except one either agreed or strongly agreed that it would be desirable to *combine* the audio technique with the visual technique.

DISCUSSION

Compared with the iPod-like visual linear menu, our audio technique has comparable speed and accuracy overall. Although initially slower than the visual technique, the audio technique was significantly faster after only 30 minutes of practice. From the outset, our audio technique was on average only half a second slower for one-level menu selections and one second slower for two-level menu selections than the visual technique.

Our analysis of the experimental logs indicated that the transition from novice to expert usage was spontaneous. Although the learning rate differed between participants, most participants were able to learn the expert behavior (tapping) quickly.

As with all experiments, the design of this study was a trade-off between a number of objectives. In our experiment, earPod differs from the iPod-like linear menu in two dimensions: using audio feedback vs. visual feedback, and using direct access to menu items vs. linear access to menu items. This second dimension (direct vs. linear menu access) in the experimental design creates a confounding factor. We nonetheless chose this approach to show that a well-designed,

expert-friendly auditory menu can be comparable to (or even beat) a widely accepted visual menu technique, at least for static menus of reasonable sizes. Given its direct access capabilities and the expert users' ability to leverage it once familiar with the menu layout, it is not surprising to see earPod eventually surpassing the iPod-like visual linear menus. Of course, the visual menu could also be redesigned to allow direct access for faster performance over time, in which case it is likely that expert performance for both the visual and audio conditions would be similar. In fact, the type of feedback used has much less impact once expert performance is reached, because as in the earPod case, expert use by definition requires minimum feedback.

However, the performance of expert use is only one factor impacting the usability of a technique. Novice performance and learning curves are important factors as well. We know that fast access is even possible for IVR systems, and that participants can eventually learn the mapping from keys to menu items and bypass listening to all the menu options. It is the difficulty when used by novices and the lack of support for easing the learning curve that makes them particularly painful to use.

Additionally, expert-friendly menu techniques are typically only possible for static menus of reasonable sizes (e.g., marking menus). For longer menus or menus with dynamic content (such as song lists), one-to-one mappings between menu items and geometric attributes (such as direction or location) are no longer feasible, and users are unlikely to gain expert performance even through practice.

We are particularly excited that the earPod selection speed is only slightly slower in the very beginning, and users learned the technique quickly to approach expert performance. These results and positive user feedback suggest a favorable user experience using earPod. By using a synchronous communication model, reactive audio feedback, and intuitive input output mapping, earPod demonstrates an audio menu selection technique with acceptable novice performance, fast learning rate, and quick transition to expert usage. The combination of these factors suggests that audio menus may finally be compelling enough to supplement or even replace visual menus in a wide range of situations.

Design for the Mobile Environment

Although the earPod technique can be used in many situations, it is particularly suited to mobile use. An important lesson learned from Pirhonen et al.'s study [29] is that gestures must be robust enough to be used when moving and that simple taps are easily triggered by accident.

To test the viability of our technique in a mobile environment, we informally asked several people to perform a few selections while standing, walking, or jogging.

We found that participants had no problem performing the glide and tap gestures to select items when standing or walking. However, tapping becomes more difficult and less accurate while running, as the touchpad started shaking in the participants' hands.

To overcome this problem, we secured the touchpad by attaching it to a hand wrap using a Velcro fastener as shown in Figure 8. With this addition, participants found it much easier to perform the taps while in motion. The fact that participants are able to use our technique (with the addition of the wristband-mounted touchpad) while running gives us more confidence that the tapping selections can be made under a variety of mobile situations, especially since most mobile scenarios, such as walking or traveling on the train, are less challenging than running.

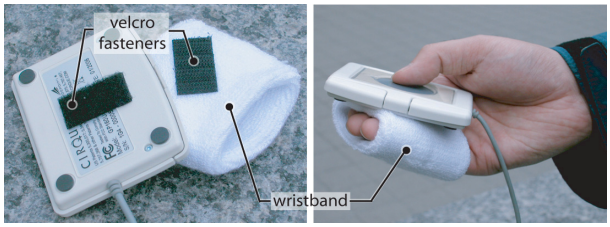

Figure 8. Using a wrist band and Velcro fasteners (left) to secure touchpad in hand (right) while the user is in motion.

Some participants in our study commented that they suspected tapping to be more likely to be triggered accidentally. This could be avoided by requiring users to hold a button or perform a squeeze gesture [16] to activate a command.

The circular touchpad could be embedded into devices of many different form factors or it could be implemented as a separate component, like a wireless remote control. Touchpads are typically very light, allowing a touchpad remote to be carried easily on a neck lanyard or a key chain. While some custom design will undoubtedly be needed to adapt earPod to such individual devices, the effort required to execute the designs should be relatively straightforward.

Dual Channel Menu

Attempts have been made to allow interfaces to be independently accessed from different sensory channels. Pauws et al. [28] propose the use of an auditory interaction technique with their music player and optional use of a visual display. Based on feedback received during our experiment, we propose using both the audio and visual modalities concurrently. In reality, either visual or audio techniques alone will often not be satisfactory. A single person may prefer one modality over the other under different scenarios. Instead of making the decisions for the user, both visual and audio techniques may be provided and users can then decide which one to use according to their specific needs.

For ease of learning, it is preferable that both techniques use a consistent input mechanism. Figure 9 presents two possible menu designs that match earPod's interaction model. In the first approach (Figure 9a), the linear order of items is preserved, so the item labels can be conveniently laid out. A shortcut icon is presented besides the label to encourage direct tapping. The second approach (Figure 9b) chooses a visual layout that matches the shape of the input device. Its layout visually reinforces the circular mental model, but it can be difficult to layout longer labels on the screen. In addition to the visual aid, the synchronized spatialized audio feedback also teaches users the absolute position of items.

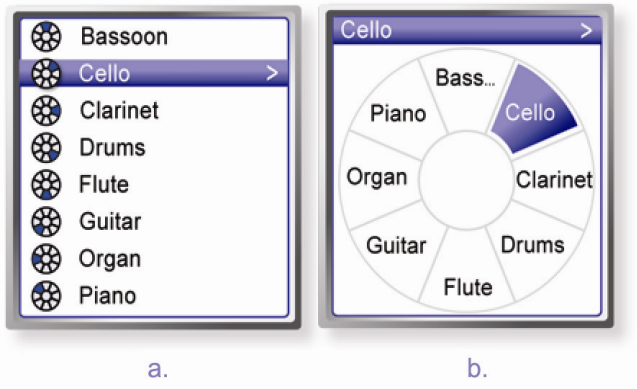

Figure 9. Two possible visual menu designs for earPod enabled devices.

CONCLUSIONS AND FUTURE WORK

We presented earPod, a touch-based menu technique with reactive auditory feedback. Our user study indicates that earPod is efficient to use and relatively easy to learn. For static menus of reasonable sizes, earPod is comparable in both speed and accuracy with an iPod-like visual menu selection technique. Although initially slower, earPod outperformed the visual technique within 30 minutes of training. This makes earPod a promising eyes-free menu technique for general use.

There are several directions for future research. In our experiment, the audio content was recorded at the speed of normal human speech. It is known that compression can speed up audio playback up to 2 times without losing clarity [3]. Future studies should explore the effect of such audio compression techniques on the effectiveness of earPod .

Furthermore, Ranjan et al. [30] found that participants were able to search audio content faster if two different audio tracks were presented simultaneously to each ear. Although our preliminary tests showed that overlapping audio signals might confuse users, this limitation might be overcome with additional design and research.

earPod currently uses simple stereo audio based on time and intensity differences between the ears. In certain environments, such as cars, use of more sophisticated and higher quality spatial audio may be feasible.

Guo and Guo's [15] experiment on Drosophila found that multi-channel training led to a better learning effect that either channel alone. An interesting area of future research would be to explore if the simultaneous use of the audio and visual channels could facilitate learning of menu structures.

In our study, we evaluated earPod against an iPod-like linear menu within the context of a simulated desktop environment. While there is no consensus on how to evaluate mobile interaction techniques, various attempts have been made to perform formal investigations beyond the stationary simulation [25, 29]. Although our informal usability testing suggests earPod is effective in motion, systematic investigation under various mobile scenarios, and direct comparison with iPod's ClickWheel in true mobile settings will be a promising future direction to explore.

We showed earPod to be effective for static menu hierarchies of reasonable size. Strategies for traversing dynamic menus and menus of arbitrary length, such as a song list are worth exploring in future studies.

With the introduction of touchpads in some recent cell phones, it seems worth exploring whether the design principles behind earPod can be used to help users navigate phone-based menus, such as IVR systems.

Finally, menu selection is only one of many possible tasks in an auditory interface; it was chosen as the first topic for this research since command selection is a fundamental building block for more complex applications. However, development of an all-encompassing interface for eyes-free operations on auditory devices is a task for future research.

ACKNOWLEDGMENTS

We thank Mike McGuffin for early discussion and feedback, Jingnan Yang for overall support, Heather Thorne, Alvin Chin, Derek Nowrouzezahrai, Yongchang Zhao for proofreading the paper, Lisa Min, Andrew Chignell, Paula Gilverson, Michelle Qu, Andrew McCutman, Noah Lockwood, John Chattoe, Annie Xu, Silvia Gonzaloz, Anna Malandrino for help running the experiment, and anonymous reviewers for their constructive advice and valuable feedback.

REFERENCES

1. KidsClick! web search for kids by librarians: http://sunsite.berkeley.edu/KidsClick!/.
2. Wikipedia: http://www.wikipedia.org/.
3. Barry, A. (1997). SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction,* 4(1). p. 3-38.
4. Bernstein, L. (1997). Detection and discrimination of interaural disparities: Modern earphone-based studies, in *Binaural and Spatial Hearing in Real and Virtual Environments*, R.H. Gilkey et. al. Editors. Lawrence Erlbaum. p. 117-138.
5. Berry, G. (1997). The Esterel Primer. *Technical report, Ecole des Mines de Paris and INRIA*.
6. Blattner, M., Sumikawa, D., and Greenberg, R. (1989). Earcons and icons: Their structure and common design principles. *Human Computer Interaction,* 4(1). p. 11-44.
7. Brewster, S., Lumsden, J., Bell, M., Hall, M., and Tasker, S. (2003). Multimodal 'eyes-free' interaction techniques for wearable devices. *ACM CHI Conference on Human Factors in Computing Systems.* p. 473-480.
8. Brewster, S. (1998). Using nonspeech sounds to provide navigation cues. *ACM Transactions on Computer-Human Interaction*. 5(3). p. 224-259.
9. Brewster, S., and Cryer, P. (1999). Maximising Screen Space on Mobile Computing Devices. *ACM CHI Conference on Human Factors in Computing Systems (Extended Abstracts)*.p. 224-225.
10. Brewster, S., Wright, P., and Edwards, A. (1993). An evaluation of earcons for use in auditory human-computer interfaces. *ACM CHI Conference on Human Factors in Computing Systems.* p. 222-227.
11. Callahan, J., Hopkins, D., Weiser, M. and Shneiderman, B. (1988). An empirical comparison of pie vs. linear menus. *ACM*

CHI Conference on Human Factors in Computing Systems. p. 95-100.

12. Doel, K., and Pai, D. (2001). A Java audio synthesis system for programmers. *International Conference on Auditory Display*.
13. Gaver, W. (1989). The Sonic Finder: An interface that uses auditory icons. *Human Computer Interaction,* 4(1). p. 67-94.
14. Gaver, W., and Smith, R. (1991). *Auditory icons in large-scale collaborative environments. ACM SIGCHI Bulletin*, 23(1). p. 96.
15. Guo, J., and Guo, A. (2005). Crossmodal interaction between olfactory and visual learning in Drosophila. *Science*, 309. p. 307-310.
16. Harrison, B., Fishkin, K., Gujar, A., Mochon, C., and Want, R. (1998). *Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces*. *ACM CHI Conference on Human Factors in Computing Systems*. p. 17-24.
17. Kurtenbach, G. (1993). *The design and evaluation of marking menus*. Ph.D. Thesis, University of Toronto.
18. Liao, C., Guimbretiere, F., and Loeckenhoff, C. (2006). Pen-top feedback for paper-based interfaces. *ACM UIST Symposium on User Interface Software and Technology*. p. 201-210.
19. Luk, J., Pasquero, J., Little, J., MacLean, K., Levesque, V., Hayward, V. (2006). A role for haptics in mobile interaction: initial design using a handheld tactile display prototype. *ACM CHI Conference on Human Factors in Computing Systems*. p. 171-180.
20. Marics, M., and Engelbeck, G. (1997). Designing voice menu applications for telephones, in *Handbook of Human-Computer Interaction*, Helander, M., Landauer, T., and Prabhu, P. Editors. Elsevier. p. 1085-1102.
21. Miller, D. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Human Factors Society Conference.* p. 296-300.
22. Minoru, K., Chris, S. (1997). Dynamic Soundscape: mapping time to space for audio browsing. *ACM CHI Conference on Human Factors in Computing Systems*. p. 194-201.
23. Mynatt, E. (1995). Transforming graphical interfaces into auditory interfaces. *ACM CHI Conference Companion on Human Factors in Computing Systems.* p. 67-68.
24. Norman, K. (1991). *The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface.* Ablex Publishing Corporation.
25. Oulasvirta, A., Tamminen, S., Roto, V., and Kuorelahti, J. (2005). Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. *ACM CHI Conference on Human Factors in Computing Systems*. p. 919-928.
26. Paap, K., and Cooke, N. (1997). Designing menus, in *Handbook of Human-Computer Interaction*, Helander, M., Landauer, T., and Prabhu, P. Editors. Elsevier. p. 533-572.
27. Pascoe, J., Ryan, N., and Morse, D. (2000). Using while moving: HCI issues in fieldwork environments. *ACM Transactions on Computer-Human Interaction*, 7(3). p. 417-437.
28. Pauws, S., Bouwhuis, D., and Eggen, B. (2000). Programming and enjoying music with your eyes closed. *ACM CHI Conference on Human Factors in Computing Systems*. p. 376-383.
29. Pirhonen, A., Brewster, S., and Holguin, C. (2002). Gestural and audio metaphors as a means of control for mobile devices. *ACM CHI Conference on Human Factors in Computing Systems*. p. 291-298.
30. Ranjan, A., Balakrishnan, R., and Chignell, M. (2006). Searching in audio: the utility of transcripts, dichotic presentation, and time-compression. *ACM CHI Conference on Human Factors in Computing Systems*. p. 721-730.
31. Resnick, P., and Virzi, R. (1992). Skip and scan: cleaning up telephone interface. *ACM CHI Conference on Human Factors in Computing Systems*. p. 419-426.
32. Roberts, T., and Engelbeck, G. (1989). The effects of device technology on the usability of advanced telephone functions. *ACM CHI Conference on Human Factors in Computing Systems.* p. 331-337.
33. Rosenberger, J., and Gasco, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. in *Understanding robust and exploratory data analysis,* Hoagan, D., Mosteller, F., and Tukey, J., Editors. John Wiley: New York.
34. Roy, D. and Schmandt, C. (1996). NewsComm: a hand-held interface for interactive access to structured audio. *ACM CHI Conference on Human Factors in Computing Systems.* p. 173-180.
35. Sawhney, N., and Schmandt, C. (2000). Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions Computer-Human Interactaction*, 7(3). p. 353-383.
36. Schmandt, C. (1998). Audio hallway: a virtual acoustic environment for browsing. *ACM UIST Symposium on User Interface Software and Technology*. p. 163-170.
37. Schmandt, C., Lee, K., Kim, J., and Ackerman, M. (2004). Impromptu: managing networked audio applications for mobile users. *ACM MobiSYS Conference on Mobile Systems, Applications, and Services*. p. 59-69.
38. Shneiderman, B. (2004). *Design the user interface: strategies for effective human-computer-interaction*. Addison-Wesley.
39. Stifelman, L., Arons, B., and Schmandt, C. (2001). The audio notebook: paper and pen interaction with structured speech. *ACM CHI Conference on Human Factors in Computing Systems.* p. 182-189.
40. Stifelman, L., Arons, B., Schmandt, C., and Hulteen, E. (1993) VoiceNotes: a speech interface for a hand-held voice notetaker. *ACM CHI Conference on Human Factors in Computing Systems*. p. 179-186.
41. Suhm, B., Freeman, B., and Getty, B. (2001). Curing the menu blues in touch-tone voice interfaces. *ACM CHI Conference on Human Factors in Computing Systems (Extended Abstracts)*. p. 131-132.
42. Witten, I., Cleary, J., and Greenberg, S. (1984). On frequency-based menu-splitting algorithms. *International Journal of Man-Machine Studies*, 21(2). p. 135-148.
43. Yin, M., and Zhai, S. (2006). The benefits of augmenting telephone voice menu navigation with visual browsing and search. *ACM CHI Conference on Human Factors in Computing Systems.* p. 319-328.
44. Zhao, S., Agrawala, M., and Hinckley, K. (2006). Zone and polygon menus: using relative position to increase the breadth of multi-stroke marking menus. *ACM CHI Conference on Human Factors in Computing Systems.* p. 1077-1086.
45. Zhao, S., and Balakrishnan, R. (2004). Simple vs. compound mark hierarchical marking menus. *ACM UIST Symposium on User Interface Software and Technology*. p. 33-42.