ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 98 (2012) 945 – 956

DOI 10.3813/AAA.918578

Combining Auditory and Visual Menus

Raine Kajastila, Tapio Lokki
Aalto University School of Science, Department of Media Technology, P.O. Box 15500, 00076 Aalto, Finland.
raine.kajastila@aalto.fi

Summary
Despite growing interest in touch screen and gesture interfaces for auditory menus, there usually is lacking interoperability between visual and auditory menus. The same control logic for both visual and auditory domains could facilitate switching to eyes-free use when needed and improve accessibility for visually impaired users. This paper presents an efficient control interface for both domains and usability tests with three interaction methods. Results show that auditory and visual menus with the same control logic can provide a fast and usable interface to control devices. Furthermore, the same auditory menu can be accessed with a gesture interface. Overall, the touch screen interaction with a visual display was fastest, the touch screen interaction with auditory display was almost as fast, while the gesture interface with an auditory display was slowest. The novel interface paradigm is explained by an example application that allows eyes-free touch screen and gesture access to a music collection on a mobile phone.

PACS no. 43.60.Dh, 43.60.Qv

1. Introduction

Research on eyes-free and audio-only interfaces has been gaining interest in recent years. These interfaces can improve the usability of a system when eyes-free operation is necessary [1]. Previous research has focused on the benefits of auditory menus in cases such as the absence or limitations of a visual display, user disability, competition of visual attention, or reduction of battery life [2]. One important, but often neglected aspect is the interoperability of the visual and auditory menus. Many benefits can be obtained by integrating these two modalities into one interface, even if they are used separately. Visual menus can be designed to be easily accessible eyes-free by using audio feedback only. The user can learn the logic while using a visual menu and when eyes-free operation is needed, an auditory menu is already familiar and can be used immediately. This is particularly useful while driving and when the visual attention should be focused on the road. Audio only menus can also be considered satisfying to use [3] and just enhancing visual menus with auditory cues can improve driving performance while using devices [4]. In addition, a touch screen can sometimes be a barrier for visually impaired users [5], but the presented design could make touch screen devices accessible to visually impaired users.

An audio interface can be more effective than its visual counterparts [2] and even the big original equipment manufacturers have introduced devices with audio-only interface. For example, Apple has introduced the iPod shuffle [6], which gives feedback to users using synthesized speech. A device, such as the iPod shuffle, without a visual display is inexpensive to manufacture and has low energy consumption. The earlier work of the authors has introduced interaction and browsing techniques that enable more sophisticated control of devices such as the iPod shuffle [7]. However, more work is needed before browsing of auditory menus, particularly when their content is unfamiliar, would be intuitive, easy and fast.

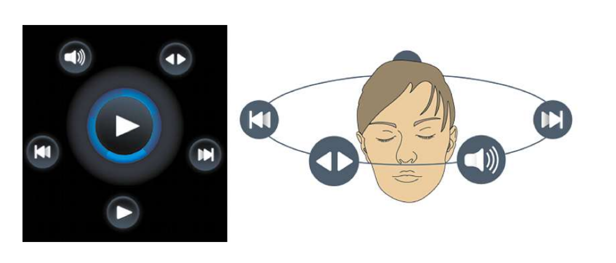

Figure 1. Circular menus can be used in both visual and auditory domains. Left: Visual menu icons on a screen. Right: The same auditory menu items are positioned around the head using spatial sound reproduction. The sounds are heard from the direction that the user points the device or touches the screen.

This article continues the earlier work of the authors [7, 8, 9] in the field of auditory menus and explores the possibility of joining visual and auditory menus (see Figure 1) by using a music application called the Funkyplayer as an example. It is the first real application that utilizes novel techniques to create an effective eyes-free user interface. Although the interface and menus are designed in the terms of audio, it is also designed to be pleasing

Received 13 February 2012,
accepted 26 August 2012.

and usable with visual feedback and offers functionality similar to common visual interfaces. Furthermore, the presented auditory menus enable efficient eyes-free browsing of hundreds of menu items [8]. It is an example that illustrates how complex tasks can be performed with auditory menus.

This article presents a functional control interface that uses circular menus with the same control logic both in visual and auditory domains. The interface is explained through novel features implemented in the Funkyplayer application. The user experiment and feedback show that the introduced touch screen interface is fast and usable in both auditory and visual domains.

2. Related work

This section gives an overview of related research, focusing on auditory menus controlled by touch screen or hand gestures and auditory user interfaces in assistive technologies.

2.1. Auditory user interfaces

Auditory interfaces have been approached in many ways. Here, audio is the main means of communication to the user, and the input methods range from normal keypads and touch interfaces to gestural interfaces.

Pirhonen *et al.* [10] tested a prototype of an eyes-free touch interface for a simple music player, in which music playing was controlled with finger sweeps on the screen. The finger sweeps from left to right, top to bottom and vice versa were used to control the volume and change the music track. Tapping of the screen was used to start and stop the track. Their study pointed out that immediate audio feedback is vital for user confidence and the interface was proven to be effective in eyes-free situations.

Savidis *et al.* [11] used the concept of auditory windows where a subset of four sound objects was simultaneously played in a spatially larger area, while others were suppressed closer together. They used a pointing interaction with a data glove, a head tracker and voice recognition to control a modifiable circular audio environment reproduced over headphones.

Egocentric circular auditory menus have been extensively studied. Brewster *et al.* [1] used a directional head nodding interface to study four simultaneously playing menu items located around the user. They found egocentric menu designs better than exocentric. Circular auditory menu structures have also been applied in Nomadic Radio by Sawhney and Schmandt [12], and in a calendar application by Walker *et al.* [13]. Visual circular menus also outperform standard pull down menus [14] and are widely used in the user interfaces of computer programs.

The study of Marentakis and Brewster [15] on audio target acquisition in the horizontal plane concluded that pointing interaction with spatial sound is successful when the user is walking. They also suggested that audio elements with feedback from egocentric audio displays could produce efficient designs.

The usability studies with touch input and a circular touchpad by Zhao *et al.* [2] showed that an auditory menu can outperform a typical visual menu used in the iPod-like devices. Their Earpod interface combined many useful features from previous research such as: 1) instant reactivity to touch input that gives control to the user without waiting periods, 2) interruptibility of the audio, where only one sound is played at a time, but its playing can be interrupted if the user chooses to continue browsing, and 3) menu items which can be accessed directly without browsing through all items.

The Foogue concept by Dicke *et al.* [16] is an example of an eyes-free interface with gesture input that does not require visual attention. Foogue can be used to control a mobile device in two modes: Menu mode and Listening mode. Menu mode is for browsing and controlling a file system that is presented with spatial sound in front of the user. In Listening mode, music, phone calls, and auditory notifications can be heard simultaneously and positioned around the head of the user. If fully implemented, Foogue would allow eyes-free control of a mobile phone and it is possible to complement it with a visual interface.

Speech recognition as an input method is also gaining in popularity, in particular after the introduction of commercial products: Voice Actions for Android [17] and Siri on iPhone [18]. With speech recognition, the voice can be used to command a mobile phone to do specific actions. Eyes-free speech recognition interfaces are mainly command oriented and, for example, eyes-free browsing for a long list of artists and selecting a song that fits your mood can be harder with speech recognition. Speech recognition is still inaccurate mainly because of language and dialect barriers, and can also be unusable in noisy environments. Furthermore, people want to maintain their privacy and prefer not to talk to their phone in public. For the above mentioned reasons, speech recognition is outside the scope of this article.

2.2. Assistive technology

Touch screens in mobile phones, home appliances and public facilities can create difficulties for visually impaired users. One of the main problems is that the visually impaired users cannot efficiently locate the graphical user interface elements on a flat surface [19]. The voiceover screen reader of Macintosh computers (OSX) and on the iPhone (iOS) can make touch screen interfaces accessible to visually impaired users. Still, touch screens are primarily designed for persons with normal vision and the use of voiceover might not be the most efficient solution. The interfaces can be designed also in terms of audio and it is also justified to implement completely different interfaces for sighted and visually impaired which engage different sensory modalities [20].

Guerreiro *et al.* [5] have implemented a gesture-based text entry method for touch screen devices. In their NaviTouch interface, all letters are accessed through vowels. The user first slides his finger vertically to find vowels that are read out loud. After hearing any of the vowels (e.g. A), the user can slide his finger horizontally to find consonants that are after that particular vowel in the alphabets (e.g. B or C). The user makes one L-shaped gesture for each successful consonant selection.

Kane *et al.* [21] used a similar L-shaped touch-gesture for browsing music tracks. In the reported experiment, ten album names were placed vertically in a list. Each item on the list could be listened to one at a time. The user first found the desired album with a vertical finger-swipe, and continued the finger movement to the right to hear the track names. Although the songs can be accessed by using only one continuous touch-gesture, it does not solve the problem when a list holds hundreds of items.

No-Look notes introduced by Bonner *et al.* [22] used multitouch text entry with the aid of a circular pie menu which was shown to be much better than using a QWERTY button arrangement with the iPhone's built in voice over. Bonner *et al.* suggested that a successful eyes-free text entry system needs to incorporate: 1) robust entry technique, 2) familiar layout, and 3) painless exploration. The same design principles can be applied to browsing eyes-free auditory menus.

Kane *et al.* [23] also studied how gestures differ between sighted and blind people to understand better how to build touch screen interfaces that work equally well for blind and sighted people. Blind people may prefer different gestures and they also may perform them differently than sighted people. Kane *et al.* reached the same conclusion as Bonner *et al.* that it is important to use familiar layouts, as well as robust gestures that reduce the demand for location accuracy.

Text entry can also be implemented using a different touch screen gesture for each character. Tinwala and MacKenzie [24] used gestures that resemble letters as input and auditory and tactile feedback to guide eyes-free entry. Letters were entered one at a time and word-level error recognition with a dictionary was used to improve accuracy. Tinwala and MacKenzie suggested that changing the speech feedback from the character-level to the word-level speeds up writing and lessens user frustration. The method was evaluated to be reasonably fast and accurate.

In the auditory menu of Tinwala and MacKenzie the word suggestions were spoken with 0.6 - second breaks and the user could pick the correct one. Due to good error correction most of the words suggested to the users were either in first or second position. Speaking the menu items one by one is the traditional way to implement menu navigation, and is largely used in Interactive Voice Response (IVR) systems in telecommunications. However, because of slowness and lack of user control it is frustrating in active use [25, 2].

3. Funkyplayer

The Funkyplayer (shown in Figure 2) is a program that is used to control the music library of an iPod Touch or iPhone. The Funkyplayer was built to demonstrate the possibilities of auditory menus, especially that audio feedback

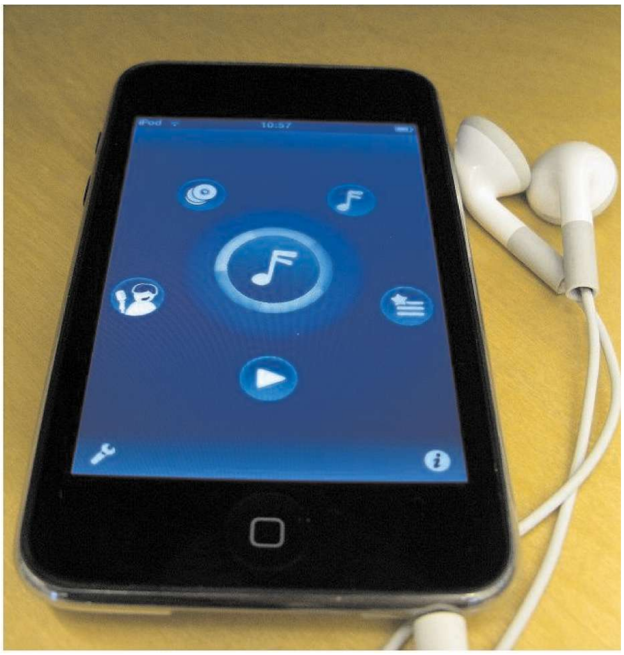

Figure 2. Funkyplayer software running on an iPod Touch.

can be efficient in menu browsing and suitable for mobile devices. The music player is only one example and many applications could benefit from design concepts that can be used without looking at them.

The design of the Funkyplayer builds upon the previous work of the authors [7, 8, 9] and the related work described in Section 2. Especially, it incorporates features highlighted by Zhao *et al.* [2] such as: instant reactivity, interruptibility of the audio and direct and fast access to the menu items. Also the three design principles of Bonner *et al.* [22] are applied to auditory menus. The menus are browsed with a simple and robust circular motion and selection is made when hearing the desired menu item. In addition, alphabetization eases the use of large menus [8]. The Funkyplayer uses circular menus, which have been found efficient to use in the visual and auditory domains [1, 14].

3.1. Interface

All menu structures in the Funkyplayer rely on egocentric circular menus. Thus, the interface is controlled with circular gestures, that can either be made on a surface (e.g. touch screen) or with wrist gestures by holding a device in the hand (see Figure 3). These two parallel interaction methods rely on a circular interaction metaphor, in which the gesture is mapped directly to the position in the menu.

In gesture interaction, simple and intuitive wrist rotations are measured with three accelerometers [7]. The iPhone is used similarly to a joystick by tilting the device slightly and rotating it 360 degrees with a gentle wrist gesture, as shown in Figure 4. The tilt angle needed to access the menu items is only 5 degrees, allowing small wrist movements and preventing any tedious turning and twisting of the wrist. The following sections concentrate more

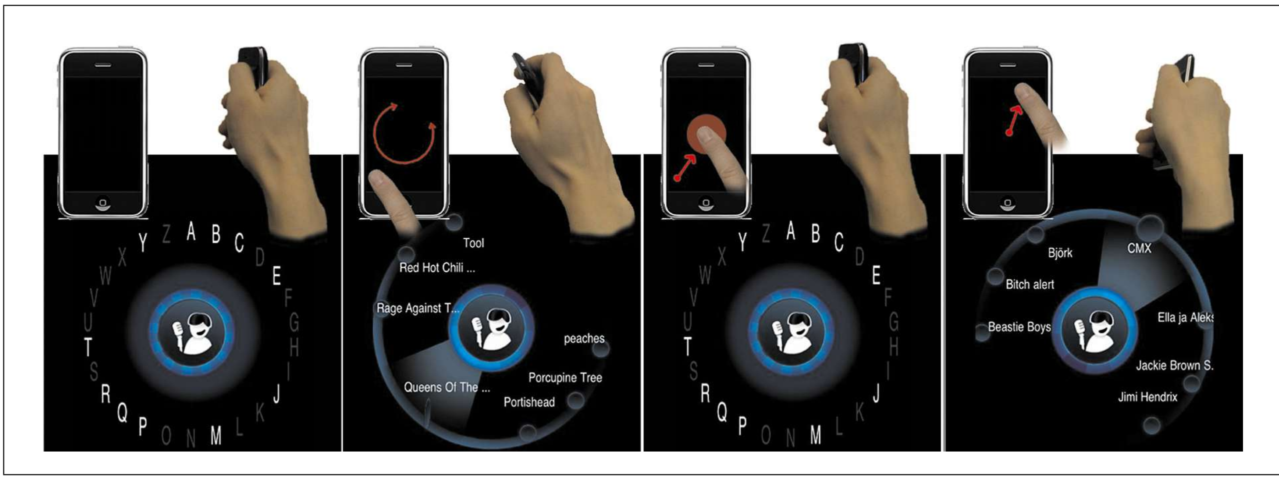

Figure 3. The screenshots from the Funkyplayer demonstrate how large number of albums can be browsed with dynamic menu item placement. The albums can be accessed by placing a finger on a sector occupying one letter. After that, all albums can be browsed with circular finger sweeps. The blue sector seen in the screenshots represent the position of the finger in the menu. The blue halo around the center is shown when the finger is in the middle of the screen or removed from the screen.

on touch screen interaction, but the same functionality is achieved by tilting and rotating the iPhone, as illustrated in Figure 3.

In touch screen interaction, the sectors extending from the center of the surface represent the menu items, as shown in Figure 5A. Menu items can be accessed directly by placing a finger on the surface, and browsing can be continued with a circular finger sweep. Selection is made by lifting the finger from the surface, which is the fastest way of making a selection. However, it can also cause wrong selections with unexperienced users or in the situation where the finger is accidentally lifted off the surface. The selection method can be altered to the needs of the user to use tap, double tap, or tapping with a second finger. The target sector of the active item expands to enable stable browsing and selecting. Note that a finger can be placed in any part of the sector, not only on the visual menu item, thus the finger position on the screen does not have to be exact during eyes-free use.

The Funkyplayer accesses the music library currently stored in the iPhone. The track names are synthesized using Flite (Festival lite) text to speech synthesis [26]. With lower quality speech synthesis, the processor of an iPhone is powerful enough for real-time synthesis and the speech feedback can be created on the fly. This saves the space needed for wav files. For higher quality speech synthesis, the names can be synthesized offline into wav files and stored for later use. Currently, only English text to speech is used, which can cause problems with names of tracks or artists in other languages.

3.2. Buttonless gesture mode

Ideally, no buttons are needed when using gestural interaction and the Funkyplayer can be controlled with a few intuitive gestures. The use of buttons can be avoided by using a quick downward motion for selecting a menu item. The motion is performed towards the gravity vector and it is not easily triggered by accident with any other movement.

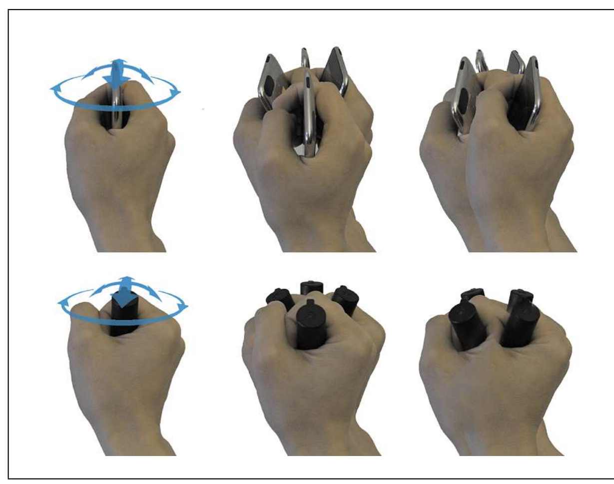

Figure 4. Wrist gestures can be used as a parallel input method to the touch screen. The gestures are performed by slightly tilting the device towards the desired menu item, as if the device was a joystick in the air. The browsing can be continued with a circular gesture. The figures illustrate the upright position (left), tilting to cardinal points (center) and half-cardinal points (right). A device shaped as a tube would fit the hand better and is more ergonomic to use (below).

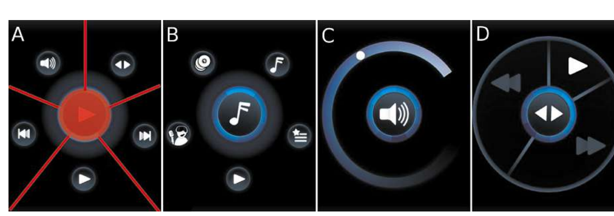

Figure 5. a) Sectors defining the menu items in the now playing menu, b) the main menu, c) the volume menu, d) the seek menu.

A locking mechanism is needed, as the menu browsing would continue if the device is just put in a pocket. The locking of the device is done by turning the device upside down for a second. Unlocking is done with the same gesture as the item selection. Auditory icons, reconstructed from two click sounds (lower and higher) played in a different order, are used to confirm the locking and unlocking. By using this feature, a user can move the device freely while listening to music without unintentional audible interruptions.

These gestures allow continuous use of the device and the user can take the iPhone out of the pocket, e.g., while biking, quickly unlock it, browse and select a new album to play, lock the device again, and put it back in the pocket. All of this can be done without looking at the device, and even while wearing a pair of thick gloves.

3.3. Spatial sound and HRTF implementation

Spatial sound is used in creating the auditory menu by positioning egocentric auditory menu items in different directions around the head of the user. The auditory menu items are indicated to the user with synthesized speech. The sounds are heard from the direction where the user points the device or touches the screen. The correlation between the direction of reproduced sound and the gesture direction can help the user associate the sound to the specific menu item location [15]. User performance can be improved with proper 3D sound design as the egocentric menu configuration becomes familiar to the user. Furthermore, spatial sound is used to separate each menu item and to make them more distinguishable when browsing with increased speed or if, e.g., music is played at the same time. Spatial sound can produce a greater sense of immersion, discovery and playfulness even in an audio space with multiple sound sources [27]. However, with multiple audio streams, spatial sound can increase cognitive load if used improperly [28].

In the Funkyplayer, the binaural implementation for headphone reproduction applies head-related transfer functions (HRTFs), which enable realistic reproduction of the spatial sound localization cues. The HRTF data were measured with the method designed by Pulkki *et al.* [29], where a loudspeaker was rotated around the subject with continuous movement in an anechoic room, and responses were measured with a swept-sine technique [30]. This process produced HRTFs every 6° in azimuth and every 15° in elevation between elevation angles of −30° and 45°. However, the circular menus of the Funkyplayer use only ear level HRTFs. The implementation of HRTF filtering uses minimum-phase HRTFs and a separate Interaural Time Difference (ITD) model. The ITD is computed with a spherical head model and minimum-phase HRTFs are modeled with 30-tap long Finite Impulse Response (FIR) filters, as presented by Savioja *et al.* [31]. The interpolation between measured positions is done separately for the ITD with fractional delays and linearly for the FIR filter coefficients. It is also possible to use the auditory menus with mono sound, although this might reduce usability.

3.4. Menu browsing

Previous studies by the authors have been made with the application and menu logic running on a computer, and devices such as the Nintendo Wiimote and iPhone were used only for controlling the computer [7, 8]. In these studies, eyes-free gestural interaction was tested by writing ten random 10-digit numbers with an auditory circular menu of 10 items. The mean time for one digit was 2.13 s with an average accuracy of 99.4% [7]. An advanced one-layer menu design was tested against two-layer menus with touch screen and gesture interaction [8]. The task was to select the correct names from a list of 156 names, as if searching a contact from a contact list. The advanced menu item spreading with hundreds of menu items in the one-layer menus was proven effective in the experiment. The Funkyplayer was created for further testing, and integrating visual menus with previously studied auditory menu browsing techniques.

As depicted in Figure 3, items in a visual circular menu can be accessed by placing a finger on the item and releasing the finger from the surface. The auditory menu works in the same way. When a finger is placed on the surface, the menu item under the finger is read out loud and it can be selected by releasing the finger. To clarify which menu item has been selected, the Funkyplayer uses a fast replay of the item mixed with a short auditory icon. A different tone of voice could have also been used, as suggested by Bonner *et al.* [22], but the advantage of the fast replay is that playback time of the feedback sound is shortened considerably, and the user can still easily recognize the content and double-check whether a correct selection was made. Auditory menus can use various types of feedback sounds, e.g., auditory icons [32], earcons [33] and spearcons [34].

Menus can be browsed by rotating the finger on the surface with circular sweeps. The menu element currently under the finger is active. In the auditory menu, the key element is immediate reactivity to user input [2]. The spoken menu items are played one by one while browsing a menu, and the user has the ability to continue to the next item, thus stopping the playback of the previous one. With slower motion, the user can hear all menu items one by one. When browsing faster, the user hears only thebeginning of the sounds. The short sounds (or spindexes [35]) represent the first letter(s) of the names and they help the user keep track of the position in a large menu. In the auditory menu described in this paper, the spindexes are automatically generated when the user browses the menu. This is achieved by instant reactivity of the auditory menu by using fast text to speech synthesis or prerecorded names. When slowing down the browsing speed, the length of the spindex is automatically adjusted and enables an efficient search method for menu items starting with the same letter, letters, or even word.

Advancing in the menu hierarchy is done by selecting a menu item from the circle and reversing is done by releasing the finger in the center of the screen. This design was chosen for consistency and having a “back” menu item always present would also occupy space from the circle. The center of the screen is easy to find during eyes-free use, because the circular gesture goes around it and the user is constantly aware of its position. When the finger reaches the center of the screen, the name of the current menu level is read out loud and mixed with a short "bubble pop"-like auditory icon indicating that the finger now is in the middle. This makes it possible for the user to always query the location in the menu structure as Kane et. al. [21] also has suggested. After a short delay, the name of the higher menu level is read out loud and the user can traverse in the menu structure.

Prerecorded info is read out loud if the user rests his finger on a menu item for a longer period of time. This way helpful information is always close if the user is uncertain about what the menu item does.

A feature especially designed for eyes-free browsing of the music collection is the possibility to play samples of the music. This is done by fading the music in after the beginning of the synthesized name of a song or an album. The music is faded out immediately when browsing to the next menu item. This feature could be used to attach other additional information to the menu elements, which could provide cues about location in the menu hierarchy, contents of the menu, or availability of the menu item.

When the volume menu item is selected, the user can adjust the volume with a circular slider by using again the same circular finger motion and releasing the finger to accept the change (see Figure 5C). The design emphasizes safety so that the volume can not be accidentally turned to the maximum level. A user who is unfamiliar with the menu might start the exploration from any part of the screen. The volume adjustment is done relative to the position where the finger is placed and not to a fixed position on the screen. Furthermore, when the volume is lowered with counter-clockwise motion it is not possible to jump accidentally to maximum volume. Instead, the end of the volume slider follows the gesture until it stops.

3.4.1. Pre-defined menu item placement

The Funkyplayer combines two different layouts for the menus: 1) Pre-defined menu item placement for a small number of elements and 2) dynamic menu item placement and spreading for a larger number of items. Pre-defined menu items are used in menus that only have a few items, whose positions can be memorized and used quickly. An alternative solution would be to always position the menu items, for example, in alphabetic order which would create consistency. However, the menu items for sound control such as "next" and "previous" are more naturally positioned to the left and right.

Pre-defined menu item placement is used in two cases: in the main menu, and in the now playing menu, as shown in Figures 5A and 5B. For example, in the now playing menu a) play/pause is at 6 o'clock (back), b) next and previous song are logically close to 9 and 3 o'clock (left and right) c) volume is at 11 o'clock, and d) seek is at 1 o'clock.

3.4.2. Dynamic menu item placement

The sub-menus such as albums, artists and songs can be browsed with one continuous touch gesture, which was

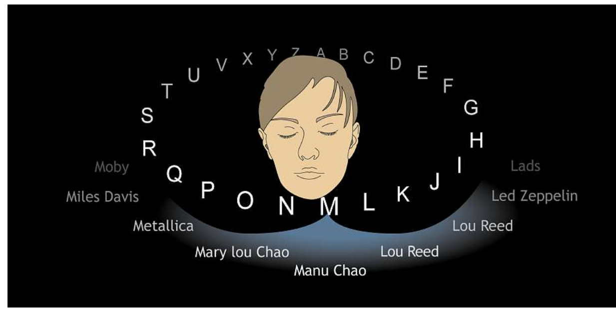

Figure 6. Menu items are positioned around the head and dynamic menu item placement allows effective browsing of large menus.

found beneficial by Kane *et al.* [21]. Furthermore, this menu design enables access to hundreds or even thousands of menu items. Until recently, it has been a challenge to browse such large auditory menus, but dynamic menu item placement been proven an efficient method, and faster than the more common two-layer solution [8].

The access to all menu items is suitable for visual and auditory menus, and utilizes alphabetization to make the menu layout familiar [22]. First, the user points towards the first letter, which is in its absolute position defined by alphabetical order, as illustrated in Figure 6. Then immediately without selection, the large list is dynamically zoomed in and the list can be browsed in alphabetical order. For example, when picking the letter M, the first menu item starting with letter M is immediately heard and seen. When browsing clockwise, next menu items can be accessed and with counter clockwise browsing the previous items are found.

The user can always go back to the first stage by sliding the finger to the center of the screen. This browsing method combines the benefit of a-priori known item positions in a static menu with large menus. Sometimes it is desirable to display the items in numerical order. This option could be used for tracks when they are displayed in an "Album's songs" sub-menu.

3.5. Application statistics

To receive feedback from the users, the first version of the Funkyplayer was available in the Apple App Store for free. The goal was to gather analytics from extended use (over one year) of the application. The application was available in Apple App Store for 5 months and was downloaded 1373 times. No advertisement was done, but people in our own university were encouraged to download, test it, and give feedback. Unfortunately, not many people gave feedback with the feedback form found inside the application. More information about the usage patterns was recorded with online mobile application analysis software, which logged the usage statistics, as permitted by the iTunes end user agreement (EULA).

The statistics show that 81% of the users were males and 19% females. Out of 1373 unique users, 217 returned to the application several times, thus most of the users tested it only once. However, after a period of one year (April

2011) there were still 2% of the users who actively used the demo application, suggesting that some people find the application useful. When starting the application, the default mode was set to touch input, but still 37.6% of the sessions were started with the gesture mode. This tells that the gesture mode also was preferred by users who obviously are used to touch screen interaction.

The total usage time of the application by all users after the launch has been 1940 hours (about 80 days). The session lengths varied a lot and most of the use sessions (46.4%) were under 1 minute in length, which does not leave enough time to really learn or operate the application. The reason for many short sessions remains unclear, although it is possible to select a song, exit the application quickly and continue listening to music in less than 30 seconds. Furthermore, 29.6% of the sessions lasted between 1-3 minutes. The remaining 24.0% of the sessions were divided into the following periods: 3-10 min (16.3%), 10-30 min (3.4%) and over 30 min (4.3%). Geographically the sessions were distributed as follows: Europe 44.1%, North America 40.1%, Asia 6.7%, Oceania 5.6 %, and the rest of the world under 1.7%.

4. User experiment

A user experiment was conducted to ascertain the usability and performance of the interaction methods and menu browsing with the Funkyplayer application. The experiment was designed to find out whether the learning of the circular menu structure is possible in audio-only mode.

The test was done while seated to keep the experiment simple and with as few changing variables as possible. The suitability of the used eyes-free interaction methods has been already tested while walking [7, 8], with a similar test setup as used in [13, 1]. The circular gesture interaction has been tested while seated and walking and there was no significant difference in selection times between immobile and mobile use [7].

4.1. Participants

Twelve participants (one female and 11 males) completed the experiment. All participants had an academic background and their ages varied from 23 to 32. The participants volunteered for the experiment, and they had no previous experience of auditory interfaces nor the interaction methods used in the experiment. Additionally, all participants were right handed, 8 out of 12 regularly used touch screen devices, and all had normal vision and hearing.

4.2. Apparatus

An iPod Touch running the Funkyplayer application was used as the test device. The auditory menu was reproduced with Sennheiser HDR HD-595 headphones connected to the iPod audio output. The screen of the iPod was used to display the visual menu. Input gestures were recognized either with the touch screen or accelerometers embedded in the device.

For consistency and reproducibility, the menu item names were synthesized with “say-command”, a built-in text-to-speech software on the Apple OSX operating system. All samples started immediately in the beginning of the sound file to ensure fast responses to user actions. The test environment was an office room and the participant was seated in front of a standard LCD screen that was used for showing the tasks to the participant.

4.3. Procedure

The experiment consisted of three tasks using auditory and visual menus measuring time and accuracy. The interaction methods were touch screen with auditory menu (TA), touch screen with visual menu (TV), and gesture interaction with auditory menu (GA).

Before each task, the current interaction method was explained with a brief demonstration after which the participants could practice the interaction method. The participants could first freely browse the menu and then practice finding 4 to 6 song names until they felt confident enough to do the actual experiment. The practice time for each interaction method took less than 5 minutes.

The task was to find and select ten songs from a list of 147 song names. The task always started from the top level menu (Main menu), from which the Songs-menu was selected. The timer was started when the name to be found appeared on the LCD screen and stopped when a song was selected from the list. After the successful selection of a song the participant stopped the music playing in the Now Playing-menu and traversed back in the menu structure to the top level menu. The setup was designed not to offer a way of correcting mistakes thus in case of a wrong selection the participants were advised to proceed without any corrections. The participants were instructed to select the given names aiming at maximum speed with minimum errors.

After each task participants filled a System Usability Scale (SUS) [36] questionnaire and answered an open-ended question about negative and positive aspects of the used interaction method. The participants were instructed not to evaluate the features of the music player itself, but the used interaction method and menu in general. In addition to the SUS questionnaire, the participants filled a short questionnaire for background information and evaluation of the interaction methods.

All interfaces were used with one hand, and on the touch screen the thumb was used for browsing. The whole experiment lasted from 30 to 45 minutes. Furthermore, the participants who gave permission, were recorded with a video camera for later analysis.

4.4. Design

The experiment was a simple factorial design, in which three different interaction methods were tested. The methods were:

- *Touch screen with visual menu* (TV), the touch screen input with visual display.
- *Touch screen with auditory menu* (TA), the eyes-free touch screen input with spatial auditory display.

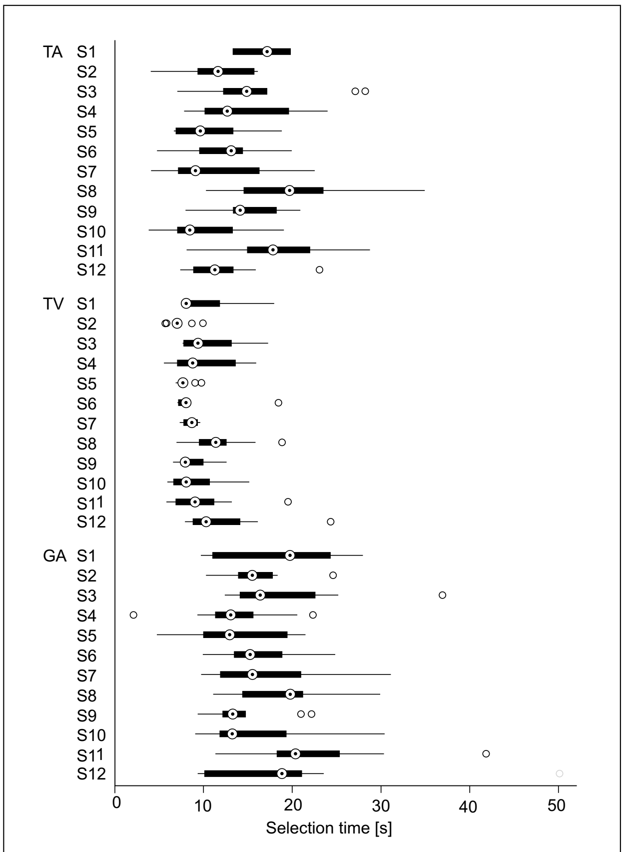

Figure 7. Time for selecting one song for each participant and each interaction method.

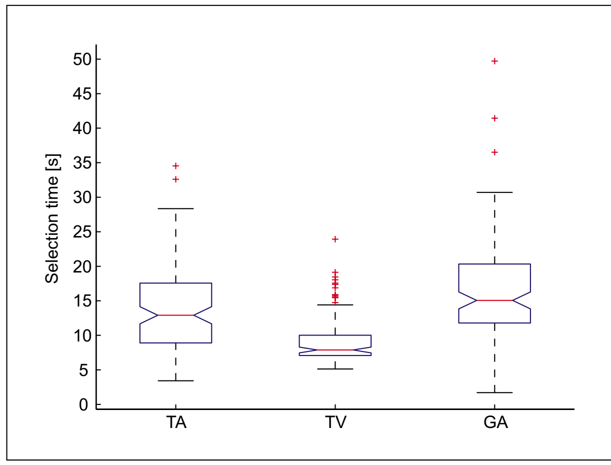

Figure 8. Time for selecting one song with three interaction methods (TA, TV, GA).

Table I. The results of the user experiment.

Time and accuracy	TA	TV	GA
Correct selections [%]	97.5	90.8	94.2
Median selection times [s]	12.91	7.87	15.06
SUS scores	TA	TV	GA
Auditory menu first	76.25	73.34	50.21
Visual menu first	75.00	60.83	41.25
Total mean SUS score	75.63	67.87	45.83

- *Gesture interaction with auditory menu* (GA), the eyes-free and accelerometer-based gesture input with spatial auditory display.

The participants were divided into two groups. The first group started the experiment with touch screen and visual menu (TV) and the second group with touch screen and auditory menu (TA). The second group did not see any visual representation of the circular menu layout before the test with TA. This was done to check if there were differences in performance when learning the menu with a different modality. The last task for both groups was the auditory menu with gestures detected with accelerometers (GA). The gesture interaction with visual menu was not included, because it is not convenient to look at the display in the hand while making the gestures. To simplify the experiment, the list of the songs remained the same for all the participants and for all tasks. However, there was not enough time to learn the list during the experiment. The list of 147 song names included 2 to 19 songs starting with each letter in the alphabet, except none starting with q, x or z.

4.5. Results

The distribution of all raw selection times was positively skewed with skewness of 1.2224, 1.3728, and 1.5030 (SE = 0.2236), for TV, TA, and GA respectively. Therefore, the median selection times of the names were compared with non-parametric one-way analysis of variance. In Figure 7, participants marked with S1-S6 started with the auditory menu (TA) and the ones marked with S7-S12 started with the visual menu (TV). No significant effect of the starting order was found in the selection times between the two groups (χ^2= 2.0037, p = 0.1569) and the individual differences were found to be large.

The median selection times of the three interaction methods are shown in Figure 8 and in Table I. The differences between rank means of tasks were also analyzed with the Kruskal-Wallis procedure. The rank means differ significantly (χ^2= 108.32, p = 0.0000). Post-hoc analysis using Tukey's least significant difference criterion ($p < 0.05$) of the three conditions showed a difference between all cases (TA, TV, GA). The percentage of correct selections for each interaction method are also shown in Table I.

Additional differential analysis (TA-TV and GA-TV individually for each subject) was performed with the Kruskal-Wallis procedure to investigate the relative changes in selection time. However, no significant effect of the starting order was found in the differential selection times between the two groups (χ^2= 0.23, p = 0.631) (TA-TV) and (χ^2= 0.03, p = 0.8728) (GA-TV). Therefore, all relative selection times without consideration of the presentation order were combined and difference between TA-TV and GA-TV was found to be significant (χ^2= 4.08, p = 0.0433).

4.6. System Usability Scale scores

At the end of each task (TV, TA, GA), participants evaluated the experience with a System Usability Scale (SUS) [36] questionnaire rating the system features according to a 5-point Likert scale. Furthermore, participants filled free form feedback about negative and positive features of the system after each task. The SUS achieved good internal reliability (Cronbach's α = 0.77 (TV), 0.79 (TA), 0.84 (GA)).

The SUS scores are summarized in Table I. A 2x3 (order x interaction method) ANOVA was conducted on the mean SUS scores and no significant effect of the starting order was found (F(1, 35)=1.73, $p = 0.2178$). Therefore, all SUS scores without consideration of the presentation order were combined and a one-way ANOVA found a significant main effect of the interaction method on SUS scores (F(2, 35)=16.42, $p = 0.0000$). Post-hoc Tukey multiple comparison of means revealed that the SUS scores of the GA differed significantly from TV ($p = 0.000$) and TA ($p = 0.001$), but TV and TA did not have a significant difference ($p = 0.2747$). To conclude, the overall SUS scores positioned the system usability between *Good* and *Excellent* (TV), *OK* and *Good* (TA),and *Poor* and *OK* (GA) [37].

5. User comments

After testing each method negative and positive feedback was collected from the participants. The following comments were obtained and grouped together in main categories.

5.1. Touch screen with visual menu (TV)

The positive feedback included 7 comments about the interface being intuitive, such as: *"Easy", "Easy to learn", "Really easy to understand"*. The menu logic and browsing also received praises: *"Navigation was generally intuitive", "The alphabetical circular menu is logical and natural", "Easy to jump to names starting with a particular letter", "Clear to see what is happening", "Really fast interface after learning the basics"*. Some participants stated the positive aspects more generally: *"Pleasant to use", "Fun and engaging"*.

The negative feedback concentrated more on details that can be quite easily improved. The most common feedback was that *"Finger blocks"* or *"Text is behind my thumb"* (6 times), *"Letters (in the alphabet) are quite close to each other"* (3 times), and *"Need to be precise when selecting"* (4 times).

The finger blocking the text can be avoided by always placing the text above the menu item. The selection area of the letters in the alphabets cannot be easily grown, but the finger position can be more intelligently approximated as it's already done in virtual keyboards, predicting the desired location more accurately. The visual sector and actually activating the menu item was misleading, because the finger (not just the sector) had to go over the menu item before it was activated. Although this known issue was accentuated to all participants during the practice, it caused the higher error rate for visual menu.

The logic and navigation also received negative comments from some participants: *"Sliding is not intuitive, but it's fault of the UI's I'm used to", "Traversing back is not intuitive to me", "Needs instructions to learn, "A bit tiring"*.

5.2. Touch screen with auditory menu (TA)

The eyes-free interaction also received 7 comments about the interface being easy, such as: *"Easy to use", "Songs were easy to find", "The experiment was easy"*. Furthermore, the word "fast" was frequently mentioned: *"Very fast to use", "Fast to move in the menu", "Fast browsing worked well"*. Four participants highlighted the eyes-free use: *"Surprisingly easy to use without looking, after you learn the application logic", "Can be used with eyes closed"*.

The eyes-free touch screen interface was also described as fun and intuitive and a few participants felt that browsing a menu with more familiar content would be even faster: *"Fast, fun and very precise", "Felt intuitive", "Would be fast with a familiar music library","The interface is exploratory and it's fun to use"*.

The negative feedback points out improvements for the system. Four participants found the selection by lifting the finger cumbersome: *"Making a selection by lifting the finger is maybe not the most convenient way", "Holding down does not feel natural, easy mistakes there", "Release to select is sometimes stressful"*. The application includes an option for using a double tap (or tap) for selection, which was not used in the experiment. In this mode the finger can always be lifted and double tapping anywhere in the screen selects the active item. Double tap (or just one tap) is a slightly slower selection method, but conveniently allows releasing the finger without making a selection.

Giving more feedback to the user would help especially when using an unfamiliar menu structure. Comments such as: *"When getting lost, knowing the current position in the menu is hard", "Where am I - functionality missing"* (2 times), and *"Recovering from getting lost needs getting used to. I'm used to visual menus"*. Even though the "Where am I" functionality was implemented, some participants did not use it or it was not clear to them. Better feedback could be implemented by attaching continuous audible information to the menu levels or items. This could provide information about location in the menu hierarchy [38].

Auditory interfaces may require users to concentrate more: *"Needs concentration", "Have to remember the order of the alphabet", "If a visual menu is not present, more thinking is needed to keep track of the position in the menu"*. The conscious effort needed to keep track of the menu position could also be reduced with proper feedback. Furthermore, three participants stated that: *"The center area is too small"*. In addition to making the center area bigger, a continuous sound instead of short "pop" could make the user more confident that the finger is actually in the center. Interestingly, three participants stated that seeing the visual menu helps in adapting to the auditory menu: *"Easier to learn after seeing the visual", "Seeing the menu during the visual test helped a lot to build a mental model"*.

5.3. Gesture interaction with auditory menu (GA)

The positive feedback in gesture control shows that some participants felt confident with the interaction method: *"Circular browsing is fluent", "The songs were quite easy to browse", "Browsing trough a list was intuitive", "Fast browsing is easy"*. Also the gesture recognition was complimented: *"The gesture detection is accurate", "The (gesture) recognition is good", "For example with a joystick style device this could be really easy to use"* Other general comments were: *"Fun", "This is novel, and I would probably use it", "Innovative", "Can be used with eyes closed"*.

The negative feedback points out some flaws in gesture detection. The device had to be held upright to access the "center button", which proved to be difficult for some participants: *"Finding the center was hard", "The center area could be bigger"* (6 times). The predefined tilt angle (5 degrees) from the upright position determines if the center menu item is active or not. This angle should be made bigger so that the center would be easier to access. However, these complaints are are also due to the unpolished selection gesture recognition. It should be implemented so that the item on which the gesture started would be selected even if the gesture ends on a different item. This would improve the accuracy and usability. Furthermore, the selection gesture received comments such as: *"Selection is not made every time", "Selection requires practice", "Selection felt funny, maybe a button would be better"* (6 times). In the experiment, the selection gesture used pre-defined thresholds and it can be improved with automatic adaptation or allowing the user to define the settings. In addition, the practice time was quite short, and every participant could learn it fast but mastering it seems to take a bit longer. Also using a "binary gesture" for selection by using a button or squeezing a (specially made) device is a more accurate method.

Some participants complained that their hand got tired: *"The hand can get tired when using longer", "A bit tiring", "Hard to keep my hand up"* (6 times). The browsing gesture was also found difficult: *"Pointing to the right direction was difficult", "Sometimes oversensitive about the gesture"*. One participant summed up the feedback well *"Really handy to be able to control the menu with wrist movement, but the implementation needs polishing"*. The slim form factor of the iPod is not optimal for this kind of circular gestures and effortless use also takes more time to learn. A device shaped like a tube or a joystick style device held with closed fingers would be more ergonomic to use with circular wrist gesture, as illustrated in Figure 4.

6. Discussion

The selection times for the three interaction methods varied significantly. The touch screen with auditory menu (TA) was on average 5 seconds slower than the visual menu (TV). However, some participants were almost as fast with TA as with TV, as seen in Figure 7. Both touch screen interaction methods (TA, TV) can be considered relatively fast and accurate to use. The SUS scores suggest the touch screen interfaces are usable and can even be fun to use. People are not accustomed to auditory and eyes-free interfaces, which may cause confusion for some users. The SUS scores for TA and TV were quite close, especially in the group that started the experiment with the auditory menu (see Table I).

The gesture interaction (GA) received the poorest results and SUS scores and it was 7 seconds slower than the visual menu (TV), which indicates that there is room for improvement. However, some participants definitely liked it and the performance of some participants was closer to touch screen (TA) performance. The speed and accuracy is still good, when taking into account that no visual feedback was given. As already mentioned, the gesture interaction (GA) would benefit from a more ergonomic control device (see Figure 4). It would also benefit from improved selection algorithm or using a different method for selecting. The hand getting tired is a problem with all gesture interfaces where the arm need to be held up for longer periods, a problem known as the "Gorilla arm effect". Thus, smaller gestures and allowing the user to keep the hand as low as possible would ease the effort needed from the muscles.

Each interaction method (TA, TV, GA) could be learned in a short tutorial and practice session which lasted less than 5 minutes. Testing the subjects after a longer usage period would probably improve the performance. Participants also wrote general comments that combining the visual and the auditory modality would improve the interface: *"The most pleasant UI would be combining both the audio and the visual UI"* , *"After using the sound the visual menu seemed deficient"*. All participants were also asked with a scale from 1 (very hard) to 5 (very easy) how easy it was to switch from visual to audio or vice versa. The average score was 3.84, which suggests that it is easy to switch between the modalities. As the input method and the logic remains the same, this kind of menu could provide easy switching between visual and auditory interfaces. One participant had a need for this kind of interface in his mobile devices and commented: *"This is excellent! I've been using visual UI's blind at times"*.

The Funkyplayer is the first effort to create systems that can be accessed easily through both visual and audio interfaces. The interface and the Funkyplayer is still a demonstration application and it can be improved in various ways. There are several features that might facilitate the use of the auditory menu. One example is attaching audible information to the menu elements, that would provide information about the location in the menu hierarchy [38]. It can also be useful to change the traditional text to speech approach, e.g., by using whispered sounds for unavailable menu items which in visual domain would be grayed out [39]. The results of the tests can be used to im-

Figure 9. Car infotainment systems would benefit from menus that are designed to be used without looking at them. A visual menu on a screen can be accompanied with an auditory menu produced with the loudspeaker system of the car. Using identical control gestures in visual and auditory menus facilitates switch between the modalities.

Figure 10. Suggestions how gesture and touch surface control with an auditory menu can be used in devices without visual display or to remotely control a mobile phone. A touch surface can be attached to a sleeve. Accelerometers attached to a glove, next to the back of the hand, can detect small wrist movements while the arm is relaxed and pointing down.

prove applications and devices that would benefit from the combination of visual and auditory menus.

7. Future applications

Auditory user interfaces can be useful when using a mobile phone in the car or when using a car infotainment systems, as illustrated in Figure 9. Mobile phones need to be used in a safe way so that the driver can concentrate on driving without visual distractions. The same applies to infotainment systems and touch screen navigators which are commonly used in cars. The interface could switch off the visual menu when the vehicle moves, and switch it on when not moving, e.g, at traffic lights. The auditory menu is used while the car is moving. The interoperability could allow seamless switching from the audiovisual menu to audio-only use.

One possibility is to embed accelerometers in a glove or a ring, which can be used to discretely and remotely control basic functionalities of a mobile phone (see Figure 10). Interacting with a smartphone without taking it out of the pocket can be useful in cold or dirty environments, e.g, while snowboarding. It is also possible to construct a small multi-functional device consisting only of internal rotation-sensing devices, e.g. accelerometers. Such a robust device without visual display can, e.g., perform all controls of a simple mobile phone. Furthermore, the interaction methods do not need to be restricted to holding physical devices. Circular freehand gestures in the air could be used and identified, e.g., with camera tracking using the front camera existing in the latest smart phones. Other use cases for camera tracking could be could controlling a public screen or interacting with a small camera in the dashboard of a car. Additionally, sophisticated hearing aids can also benefit from control methods that are effective, robust, discrete, and buttonless.

8. Conclusion

This paper introduced a menu design that can be used with the same control logic both in visual and auditory domains. Interoperable auditory and visual menus were demonstrated with the Funkyplayer application, which uses a novel browsing method for auditory menus. This paper presented an usability test with three interaction methods. Results show that the introduced menu design is easy and intuitive to learn without extensive training. Especially, the usability of the touch screen interaction methods was found good and switching between modalities easy. An auditory menu was combined with a visual menu by using synthesized speech samples. Spatial sound processed with HRTFs was used to display menu items as well as to give feedback about the menu item selection to the user. The presented ideas can be applied as an alternative control method to mobile devices, hearing aids, car infotainment systems and public touch screens.

Acknowledgment

The research leading to these results has received funding from the Finnish Funding Agency for Technology and Innovation (TEKES), the Academy of Finland, project no. [119092 and 218238] and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636]. Thanks to all who were involved with the Koala project and contributed to the building of the Funkyplayer.

References

[1] S. Brewster, J. Lumsden, M. Bell, M. Hall, S. Tasker: Multimodal 'eyes-free' interaction techniques for wearable devices. Proc. CHI, ACM, 2003, 463–480.

[2] S. Zhao, P. Dragicevic, M. Chignell, R. Balakrishnan, P. Baudisch: Earpod: Eyes-free menu selection using touch

input and reactive audio feedback. Proc. CHI, ACM, 2007, 1395–1404.

[3] J. Sodnik, C. Dicke, S. Tomaič, M. Billinghurst: A user study of auditory versus visual interfaces for use while driving. Int. J. Hum.-Comput. Stud. **66** (2008) 318–332.

[4] M. Jeon, B. K. Davison, M. A. Nees, J. Wilson, B. N. Walker: Enhanced auditory menu cues improve dual task performance and are preferred with in-vehicle technologies. Proc. AutomotiveUI '09, 2009, 91–98.

[5] T. Guerreiro, P. Lagoá, H. Nicolau, D. Gonçalves, J. Jorge: From tapping to touching: making touch screens accessible to blind users. IEEE Multimedia, 15, 2008, 48–50.

[6] iPod shuffle: portable music player. http://www.apple.com/ipodshuffle/, 2011.

[7] R. Kajastila, T. Lokki: A gesture-based and eyes-free control method for mobile devices. CHI'09 extended abstracts on Human factors in computing systems, ACM, 2009, 3559–3564.

[8] R. Kajastila, T. Lokki: Eyes-free methods for accessing large auditory menus. Proc. ICAD, 2010, 223–230.

[9] R. Kajastila, J. Vehviläinen, P. Palm, T. Lokki: Funkypod: Integrating auditory and visual menus. Proc. IWPASH, 2009.

[10] A. Pirhonen, S. Brewster, C. Holguin: Gestural and audio metaphors as a means of control for mobile devices. Proc. CHI, 2002, 291–298.

[11] A. Savidis, C. Stephanidis, A. Korte, K. Crispien, K. Fellbaum: A generic direct-manipulation 3d-auditory environment for hierarchical navigation in non-visual interaction. Proc. ASSETS, 1996, 117–123.

[12] N. Sawhney, C. Schmandt: Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. ACM Transactions on Computer-Human Interaction **7** (2000) 353–383.

[13] A. Walker, S. A. Brewster, D. McGookin, A. Ng: A diary in the sky: A spatial audio display for a mobile calendar. Proc. BCS IHM-HCI, Springer, 2001, 531–540.

[14] J. Callahan, D. Hopkins, M. Weiser, B. Shneiderman: An empirical comparison of pie vs. linear menus. Proc. CHI, 1988, 95–100.

[15] G. Marentakis, S. Brewster: Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane. Proc. CHI, 2006, 359–368.

[16] C. Dicke, K. Wolf, Y. Tal: Foogue: eyes-free interaction for smartphones. Proc. MobileHCI '10, 2010, 455–458.

[17] Voice Actions for Android. http://www.google.com/mobile/voice-actions/, 2011.

[18] Siri. http://www.apple.com/iphone/features/siri.html, 2011.

[19] K. Usui, M. Takano, Y. Fukushima, I. Yairi: The evaluation of visually impaired people's ability of defining the object location on touch-screen. Proc. ASSETS, 2010, 287–288.

[20] F. Winberg, J. Bowers: Assembling the senses: towards the design of cooperative interfaces for visually impaired users. Proc. CSCW '04, ACM, 2004, 332–341.

[21] S. K. Kane, J. P. Bigham, J. O. Wobbrock: Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. Proc. ASSETS, 2008, 73–80.

[22] M. Bonner, J. Brudvik, G. Abowd, W. K. Edwards: No-look notes: Accessible eyes-free multi-touch text entry. Proc. Pervasive '10, 2010, 409–426.

[23] S. K. Kane, J. O. Wobbrock, R. E. Ladner: Usable gestures for blind people: understanding preference and performance. Proc. CHI, 2011, 413–422.

[24] H. Tinwala, I. S. MacKenzie: Eyes-free text entry with error correction on touch screen mobile devices. Proc. NordiCHI, 2010, 511–520.

[25] M. Yin, S. Zhai: The benefits of augmenting telephone voice menu navigation with visual browsing and search. Proc. CHI, 2006, 319–328.

[26] Flite (festival-lite) text to speech synthesis engine. http://www.speech.cs.cmu.edu/flite/index.html, 2011.

[27] Y. Vazquez-Alvarez, I. Oakley, S. Brewster: Auditory display design for exploration in mobile audio-augmented reality. Personal and Ubiquitous Computing Journal (2011) 1–13.

[28] Y. Vazquez-Alvarez, S. A. Brewster: Eyes-free multitasking: The effect of cognitive load on mobile spatial audio interfaces. Proc. CHI, 2011, 2173–2176.

[29] V. Pulkki, M.-V. Laitinen, V. Sivonen: HRTF measurements with a continuously moving loudspeaker and swept sines. Presented at the 128th Convention of the Audio Engineering Society. J. Audio Eng. Soc. (Abstracts) **58** (2010) preprint 8090, 178.

[30] A. Farina: Simultaneous measurement of impulse response and distortion with a swept-sine technique. Presented at the 108th Convention of the Audio Engineering Society. J. Audio Eng. Soc. (Abstracts) **48** (2000) preprint 5093, 350.

[31] L. Savioja, J. Huopaniemi, T. Lokki, R. Väänänen: Creating interactive virtual acoustic environment. J. Audio Eng. Soc. **47** (1999) 675–705.

[32] W. Gaver: Auditory icons: Using sound in computer interfaces. Journal Human Computer Interaction **2** (1986) 167–177.

[33] M. Blattner, D. Sumikawa, R. Greenberg: Earcons and icons: Their structure and common design principles. Journal Human Computer Interaction **4** (1989) 11–44.

[34] B. N. Walker, A. Nance, J. Lindsay: Spearcons: Speech-based earcons improve navigation performance in auditory menus. Proc. ICAD, 2006, 63–68.

[35] M. Jeon, B. N. Walker: Spindex: Accelerated initial speech sounds improve navigation performance in auditory menus. Proc. HFES, 2009, 1081–1085.

[36] J. Brooke: SUS: A quick and dirty usability scale. Usability Evaluation in Industry, 1996, 189–194.

[37] A. Bangor, P. Kortum, J. Miller: Determining what individual sus scores mean: Adding an adjective rating scale. Journal of Usability Studies **4** (2009) 114–123.

[38] E. Sikström, J. Berg: Designing auditory display menu interfaces - cues for users current location in extensive menus. Presented at the 126th Convention of the Audio Engineering Society. J. Audio Eng. Soc. (Abstracts) **53** (2009) preprint 7659, 150.

[39] M. Jeon, S. Gupta, B. K. Davison, B. N. Walker: Auditory menus are not just spoken visual menus: A case study of andunavailable menu items. CHI'10 extended abstracts on Human factors in computing systems, 2010, 3319–3324.