Evaluating Mid-air List Interaction for Spatial Audio Interfaces

Christina Dicke
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin
Berlin, Germany
christina.dicke@ixds.com

Jörg Müller
Department of Computer Science, Aarhus University
Aarhus, Denmark
joerg.mueller@acm.org

ABSTRACT

Selecting items from lists is a common task in many applications. For wearable devices where no display is available, list selection can be challenging. To explore potential solutions we present four user studies evaluating mid-air gestures to interact with lists in an eyes-free interface. We found that a spatialized audio list in the shape of a 110 degree arc angled towards the dominant hand was a comfortable and usable layout for most users. A selection takes less than 10.6 seconds on average and error rates are below 4% when users locate and select an item in an unknown, unordered list of 20 items. For lists of 10 items the mean selection time is 5.5 seconds or less, and error rates drop below 1.4%. We compared monophonic to binaural playback of feedback sounds (musicons) and found no statistical difference for task completion times or error rates between the conditions. We also implemented and evaluated a music player application to showcase spatial audio list selection in an applied scenario.

Keywords

Auditory Display, Mid-air Gestures, List Selection, Direct Manipulation

Categories and Subject Descriptors

H.5.2. [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces—*Interaction styles*

1. INTRODUCTION

Selecting items from lists is a common task in many applications, for example, people use menus to navigate through options, select names from contact lists, or create and share their personal playlists. Selecting an item from a list usually involves browsing the list and then selecting one or several items. Browsing and selecting requires a representation of the list, like a visual display, and some form of user interaction, such as using a scrollbar with a mouse or a swipe gesture on a touchscreen.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SUI'15, August 8–9, 2015, Los Angeles, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3703-8/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2788940.2788945 .

Although this form of list presentation and interaction is general, and will work in almost any traditional task situation, it is not always optimal. Scenarios–in which the device is too small for a usable visual display, where the visual display has a very low resolution, or the input capabilities are limited–are poorly supported by traditional list interaction techniques. With the spread of wearable technology and the accompanying miniaturization of I/O capabilities, traditional list selection may become slow and frustrating.

A number of eyes-free solutions have been proposed to compensate for insufficient display space. Hardware buttons, headphone cable switches or small touch-sensitive areas simplify quick interactions with a small device, such as a watch or music player, but only a reduced set of discrete interactions is supported by these methods. Cord input [32] can provide continuous input through touch location, twisting, bending and pulling but is error-prone and may not always be accessible. Complex interaction can be accomplished with speech recognition [35, 30], which offers direct, hands-free user input. But speech-recognition is still suboptimal in noisy environments or in the presence of multiple speakers.

Gestural input on or with a device is an alternative [27, 41, 36, 16, 26] and touch gestures–abstract mappings of discrete commands–can extend the number and complexity of executable functions beyond a simple switch interface. However, touch gestures cannot efficiently support tasks that require direct and continuous feedback like changing scrolling speeds for navigating through and interacting with large lists. Users either have to skip through the list step-by-step or memorize a unique gesture mapped to that function, thereby increasing cognitive load. Furthermore, touch gestures still require a physical device that is touched, held or otherwise stabilized in an accessible position, and errors may arise due to aging or clogged devices, or fumbled access through clothing.

Gustafson et al [15] proposed imaginary interfaces, a free-hand spatial interaction technique that gives users direct access to an invisible display. Spatial memory can aid orientation and enable quick access to items even after an extended period of time. The drawback here is that users have to rely on their visual short-term memory and do not receive any kind of feedback. For unknown or long lists, efficient interaction is difficult to achieve without a continuous spatial representation of the list's items and state.

To address the difficulties of the limited interaction capabilities of wearable technology–and the resultant issues of scalability, efficiency and accuracy–we conducted three user studies to investigate direct spatial manipulation for list selection facilitated by an auditory display. Because the system was controlled by mid-air gestures, miniaturization of physical input and visual output capabilities did not impact the user interaction. Furthermore, taking advantage of kinesthetic and spatial memory effects quickened the interaction, making it useful for both microinteractions [4], and for

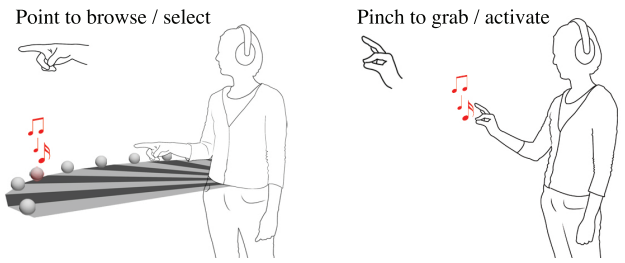

Figure 1: List selection with lists arranged in an arc centred on the user. By pointing at items users can browse the list and select items (Left). Once a song is found it can be activated through a pinching gesture (Right).

more complex tasks like copying, pasting, and multiple or ranged selection.

2. STUDIES OVERVIEW

2.1 Study 1: Physical constraints, angle, location, distance

We verified the basic physical constraints of interacting with a spatial display using mid-air gestures. We assessed the range, location, and distance of such a display, as perceived by users who varied widely in what they found comfortable. Different preferences for distance lead some users to over- or undershoot a target for which spatial positions were fixed. Thus, we thus changed the grab-to-select gesture to a point-to-select gesture. Based on our findings, for the next two studies we implemented an arc-shaped list that began at 11 o'clock and ended at 3 o'clock.

2.2 Study 2A and 2B: Impact of list length and playback type

The second study revealed the effect of list length on users' speed and accuracy. We tested users on a list of 10 items and on one double that size. Error rates were marginal even for lists with 20 items (3.8%). Mean selection times were 5.5 seconds for 10-item lists and 10.6 seconds for 20-item lists. To measure the effect of sound rendering, we compared binaurally rendered to monaurally rendered playback for 10-item lists. Both conditions showed a strong learning effect but there were no differences in task completion times or error rates. Spatial memory and the learning effect were probably facilitated by the physical pointing gesture rather than the binaural rendering. Although both conditions were equally efficient, participants tended to find the binaural condition easier to use.

2.3 Study 3: Evaluation in an applied context

For the final study we built a music player to serve as an example application for a list selection task. We compared a version of the music player that is rotationally fixed to the world against one that is user-fixed. We also explored a way to scroll through and obtain an overview of lists of 60 items and to adjust the volume by continuous horizontal movement of the hand. In general, participants liked the player and its direct and physical interaction style. All participants could successfully scroll through the list and manipulate items. Participants liked the user-centred display but preferred if the display did not rotate to follow their head.

3. RELATED WORK

Our review of previous work will focus on two main subjects: work in the general field of gestural interaction–particularly using mid-air gestures–and 3D auditory displays that utilize gestures for user interaction.

3.1 Gesture-based Interaction

In an exploratory study, Wolf et al. [39] investigated how users would spontaneously interact with a spatial auditory display. Users were given a dummy device and encouraged to perform any gesture of their choosing, on or with the device, to solve 20 typical tasks, including item selection and manipulation. Wolf et al. observed that participants created gestures through associations with other known interaction techniques or analogies from other domains. These ranged from discrete one-dimensional gestures performed on the device to continuous three-dimensional gestures and combinations thereof. For interacting with spatial auditory displays they recommend small combinable gesture sets, gesture inversions for do- and undo-commands, and preferred discreet minimalistic gestures over expressive ones.

Marentakis & Brewster [20] evaluated three different gestures for browsing and selecting in a 3D soundscape. Participants used either their head, hand or a touch tablet. Eight sound sources were positioned at a distance of 2 meters on a 360 degree ring around a user's head and could be selected by turning their head towards the source, pointing at the source or browsing with a pen on the tablet. Marentakis & Brewster found that the tablet condition was significantly more accurate than the other two techniques. They observed that a significant number of participants tried to point without turning their bodies, which influenced the accuracy of the browsing and selection process. Motion Marking Menus, proposed by Oakley and Park [25], are a gestural menu technique based on rotations of a handheld device around a single axis over a 90 degree range. A user can select items from a marking menu by tapping on a touchscreen or pressing a button. Oakley and Parker found that a menu system containing 19 commands gave optimal performance and was well suited to kinesthetic and eyes-free interaction.

Gustafson et al. [15] introduced imaginary interfaces, a free-hand spatial interaction technique that allows users to create their own imaginary interfaces. Relying on visual short-term memory and a reference frame given by users' own non-dominant hands, invisible objects can be drawn and pointed at in 2D space. In three studies, Gustafson et al. showed that participants could create and annotate simple drawings, and point at locations, without requiring any feedback. They recommended exploiting visual or kinesthetic features, such as the reference hands' finger length, to support users' memory of objects' positions.

Ashbrook et al. [3] use a finger worn ring to control up to eight choices in a menu. A user can turn the ring and by means of a magnetic field sensor the ring's rotations around the finger can be mapped to elements in a list. A selection is made by moving the ring along the finger. Although this technique has a high social acceptability the list size is restricted to 8 elements.

ShoeSense, proposed by Bailly et al. [5], is an eyes-free interaction technique for mobile devices. A shoe-mounted depth-camera is used to recognize hand-gestures, such as a radial pinch, a finger-count, or a triangle formed between the right hand and left arm. ShoeSense can be used to control an eyes-free application by mapping gestures to operations. Participants found gestures required low physical and mental demand. Although gestures had a high social acceptability in general, interviews revealed that such acceptability varied with the user's location.

3.2 Spatial Auditory Displays

Pirhonen et al. [27] developed the TouchPlayer, a hip-worn mobile music player controlled by gestures performed on the touchscreen of a PDA. One-dimensional discrete gestures executed with one finger were mapped to the player's functions, like a sweep across to skip to the next track. Compared to a standard visual interface, the TouchPlayer significantly reduced workload and task completion times without impacting error rates. However, menu navigation or item selection were not supported and the PDA had to be worn on a belt.

PocketMenu [26] was similar to TouchPlayer in that it utilized a touch-enabled device to control a music player. A limited number of menu items were laid out along the screen's border and could be selected by a swipe gesture towards the screen's center. Users received vibro-tactile and synthesized speech feedback. PocketMenu supported discrete and continuous input, e.g. for volume adjustment.

Dicke et al. [12] built a user-centred spatial sound display for navigating between multiple sounds. The auditory display consisted of three virtual rings at different distances, on which sound streams were positioned. Users could perform discrete two-dimensional gestures with a mobile phone to rotate rings, move sources between rings, or focus on a source. Dicke et al. showed that users could quickly navigate between a limited number of sources by performing flick and pan gestures with a device.

Building on [12] Dicke et al. proposed Foogue [13], a spatial auditory display concept supporting item selection and manipulation. User interaction is supported through a combination of discrete and continuous two-dimensional gestures performed with and on a touch-enabled device. Foogue has two modes, menu mode and listening mode. In menu mode users can navigate through hierarchical structures and select single or multiple items, which are to be displayed as players in listening mode. These players can be moved freely by point-move-release gestures in an egocentric, two-dimensional, 360 degrees space. Foogue offers many solutions to the challenge of designing a mobile music player, however, it remained a design concept and was never evaluated in a user study.

Kajastila & Lokki [17] compared three methods for controlling circularly and rectangularly arranged auditory and visual menus. The circular display presented twelve spatially arranged items (numbers) spread evenly on a virtual circle surrounding a user's head. Users could make a selection by either rotating their hand or moving it towards a number. Kajastila & Lokki found that free-hand gestures were fast and accurate and that smaller circular gestures were preferred over spacious circular gestures, as they reduced effort and time. By using hand-rotation in mid-air, they overcome the necessity of holding a physical device for selecting from short lists.

Müller et al. [24] developed an interactive system to "touch", grab and manipulate sounds in mid-air. They could show that users can locate, walk towards and touch spatially rendered sounds with a high accuracy and without any visual feedback.

4. CONTRIBUTION

We present four user studies exploring a direct manipulation approach utilizing mid-air gestures to interact with spatialized lists in an auditory display. In the first three studies we investigated the physical constraints of spatialized lists in order to define a physiologically adequate display angle, and we looked at the effects of list length and sound rendering on selection time and error rate. To validate this work in the context of a real application, we developed a music player controlled by mid-air gestures for the fourth study. Aside from a general evaluation of the player, we also used it to learn about participants' preferences with regard to navigating a list of 60 items. We believe some of our findings are independent of the display's modality and could generally contribute to the design of gesture-based interactions, for example in three-dimensional, immersive environments like games, or in exploring interface alternatives for visually impaired users.

5. DESIGN SPACE OF LISTS IN AUDITORY DISPLAYS

Cockburn et al. [10] systematically described the design space for gestural interaction with and without visual feedback in a framework for air pointing. Taking this work into consideration, in this section we discuss the properties that we believe are essential to the design of spatilized lists and list-selection.

5.1 Representation

Efficiently representing a list that contains more than a dozen items is a challenge in auditory display design. Due to the temporal nature of sound, playback time linearly increases with the number of items in a list, and therefore the time a listener needs to find an item. To reduce this display time, researchers have proposed several solutions. For lists of sound files the most obvious solution is not to play the file itself but a much shorter handle or abstract representation. This has been done in the form of earcons [9], spearcons [38] or musicons [21]. The benefit of these methods is the reduction of playback time while maintaining a fair degree of intelligibility. Playing files or handles synchronously or with onset intervals has been explored as an additional solution [9, 34, 14] but is limited due to masking effects.

5.2 Display dimensionality

Although lists are one-dimensional, the way they are presented to a user is not necessarily limited to one dimension. Display dimensionality is distinct from the rendering technique as it refers to the layout of the display and not to how sounds are played to users. Examples of one-dimensional displays include the iPod Nano's VoiceOver feature and Audio Bubbles [22]. Examples of 2D displays include a multi-party conference [12] and interacting with an in-vehicle menu [34]. Examples of 3D displays include accessing a music collection [31] and outdoor navigation [37] or the audio progress bar [11]

5.3 Rendering

Independent of the display's dimensionality, the sound itself can be reproduced in varying dimensions. It can be rendered binaurally — coming from a position located outside of a listener's head — , with directionality, as in stereophonic sound, or monaurally, i.e. located inside a listener's head.

5.4 Scrolling

In visual displays, there are many well-established techniques for helping users to access long or structured lists, including sectioning, pagination, hierarchization, and zooming. There are only few such methods developed for auditory displays. [31] used a strong physical metaphor in which a music collection is divided into navigable rooms. [41, 17] built auditory menus in the style of a Marking Menu to access items in a hierarchically structured list.

5.5 Overview

An overview of items in a list (or of one's position there) should be readily available and easily processable. Again, the temporal nature of sound makes this a challenge. Researchers have addressed

this, for example [8] in the context of hierarchical menus and [18] complex data sets.

5.6 Interaction

Different interaction styles and paradigms have been proposed for interacting with lists. Popular solutions are gestures performed on a device [27, 41], with a device [12, 19] or with body parts [17, 20]. Interaction can be in discrete steps as in [12] or continuously as in [19]. Depending on position and dimensionality, interaction can be limited to a range [17], an area [5] or it can be ubiquitous [12].

5.7 Location

Most mobile auditory displays play audio relative to a user's height at face level [31, 27, 20, 41, 12]. A likely reason for this is that usually headphones are used to display 3D sound, which naturally positions sources at head-height. A display could also be anchored at other body parts such as shoulders, hands or feet.

5.8 Translation

Translation refers to how a display reacts to movement. If it is fixed on a user (egocentric) it will move when the user moves. If it is fixed on a location independent of a user (world-fixed or exocentric), a user can move towards it or away from it. World-fixed displays with absolute positioning are popular for way-finding and navigational tasks [19].

5.9 Rotation

Rotation refers to how a display reacts to a user's rotational movements. For example, if a display is user-fixed and the user turns their head the display also rotates. This is often the case in displays that use headphones and do not compensate for head movements.

6. GESTURAL INTERACTION DESIGN OVERVIEW

We overcome the need for display and interaction space on a device by choosing mid-air gestures. Enabling users to mimic how they would naturally interact with physical objects draws on their already available implicit knowledge of the movement ("knowing how") and offers a mental model that is easily accessible and learnable ("knowing what") (cf. [2]). For the design of our list-selection approach we focused on optimizing gestures for ease of use and not reflecting the limitations of the current state of technology. Aspects contributing to the usability of free-hand gestures as summarized by Baudel & Beaudouin-Lafon [6] were taken into account as well as factors impacting the joy of use and social acceptability. In particular, we focused on these factors:

1. **Physiological Adequacy**: Gestures should be simple to perform, require minimal muscle stress and effort, and be designed for repetitive use.

2. **Contextual Adequacy**: Interaction should be intuitive, easily discoverable and sensible in the context of the application. Logical consistency within the gesture set should be maintained, for example by gesture reversion for do/undo-commands as recommended by [39].

3. **Social Acceptability**: Gestures should be socially acceptability to encourage adoption of gestures. Rico & Brewster [29] conducted studies on the social acceptability of device and body based gestures performed in the wild. They found that location and audience have a significant impact on users' willingness to perform gestures. Subtle imitations of everyday gestures, like shaking or tapping, were rated more acceptable in public than large or noticeable gestures, like a shoulder or nose tap. Besides social factors, the ease with which gestures can be performed had an impact on ratings. Physically uncomfortable gestures, such as head nodding, foot tapping, and wrist rotation were rated lower than easy to perform gestures. Montero et al. [23] found users' acceptance for performing a gesture in public places was influenced by whether they thought bystanders were able to interpret the intention of the gesture. Yi et al. [40] found that social respect and avoiding interruption to social activities are important user motivations for using eyes-free interaction.

Using mid-air gestures moves the interaction towards the *expressive* end of the scale defined by Reeves et al. [28] in their classification of public interfaces. Following their definition we used *suspenseful* gestures: the manipulation is obvious but as the effect is displayed aurally it is only revealed to the user wearing the headphones and not the bystanders. We chose self-explanatory gestures for our study, which are based on everyday life's physical interactions and allowed for subtle and expressive gestures to address issues of social acceptability.

7. STUDY 1: PHYSICAL CONSTRAINTS, ANGLE, LOCATION, DISTANCE

In this first exploratory study, we verified the basic spatial constraints of the interface. While its position in space is already restricted by human anatomy, i.e. by where it is comfortable to reach, we took a closer look at participants' subjective perception of where they would want the interface to be. Specifically, we addressed these research questions:

- RQ1: What is perceived as a comfortable angular window size for pointing at sources in space?
- RQ2: What is the preferred location for this window?
- RQ3: What is the preferred distance from hand to body for spatial pointing?

7.1 Experimental design

Participants were first introduced to the concept of the "point and select" interaction style of the list. They had 5 to 10 minutes to familiarize themselves with an example list of 180 degree angular range starting at -90 degrees. We started measurements once participants found their "comfort zone" and stopped measuring after four repetitions. Dependent variables for this study were *angle size*, *angle position*, and *radius*, i.e. distance from the center of the hand to the center of the head.

7.2 Task

Participants demonstrated their preferred shape and position of the list. They pinched (touched thumb and index finger) where they would want the list to start, circumscribed the range with their hand at a comfortable distance from their body, and then pinched again where they would want it to end.

7.3 Participants

We recruited 16 right-handed participants from our institutions' database (5 male). They were between 25 and 68 years old (mean age 36 years) and received a small compensation for their time.

7.4 Technical Setup

For all experiments we used an Optitrack [1] optical tracking system with 16 cameras for high-precision localization of participants' head and hand positions. These were received in a Processing [2] sketch from where sound positions were controlled. Via a Pure Data [3] patch, sound control events were routed to the Sound Scape Renderer [1] (SSR)[4] for binaural rendering. HRTFs measured in a small studio room as described in [7] were applied to make the rendering slightly echoic. Participants wore AKG K601 reference headphones with compensation filters applied and also a custom made glove. Optical markers were sewn onto the glove and a pinch of thumb and index finger was wirelessly transmitted and recognised by the Processing sketch.

7.5 Results and discussion

As shown in Fig. 2, participants varied strongly in all three aspects measured. Some were comfortable extending their arms almost completely and circumscribed nearly 180 degrees (p2, p4, and p10), but others preferred a very narrow frontal range (p11, p13, and p18). Most participants started their movement at approx. 11 o'clock (-15.39 degrees). The mean angle range was 105.32 degrees and most participants ended between 2 and 4 o'clock. We could observe a similar variance in hand-to-head distance. The shortest distance was 0.27 meters and the longest 0.66 meters with a mean of 0.55 meters.

An overlay of all 16 datasets is shown in Fig. 3. The head position is marked by the two grey lines crossing at (0.0,0.0). The black dashed line shows the overall mean angular hand-to-head distance. Given that arm lengths varied between participants and acknowledging that the average arm length of a human correlates with their height, we assumed an impact of gender on the results. As women are usually smaller and hence have a shorter arm length, we hypothesized that shorter distances were preferred by female participants. We found that this is not the case. Interestingly, both very small and close ranges and large and more distant ranges were circumscribed by female participants. Data from male participants is shown as p01, p03, p09, p16, and p17 in Fig. 2.

Concluding from the results, subjective preference seemed to be the strongest influence on the users' comfort with their chosen *angle size*, *angle position*, and *radius*. However, age or physical health may also have an influence but were not evaluated in this study. The high variance in the results seem to dispel a notion that we influenced participants by priming them with a 180 degree angle. Overall, an angle of approx. 110 degrees starting at -30 degrees and ending at 90 degrees seemed an acceptable compromise.

We took another important learning from this first study: as we saw such a strong variance in what was perceived as a comfortable hand position for pointing at objects in mid-air, we changed the initial design from a "touch the source to select" style to a "point at the source to select" style. While in the initial approach the hand position (h_x,h_y) had to be in a certain radius around a source's position (s_x,s_y) to initiate a selection, the new approach registers a selection based on whether the hands' position is in a isosceles triangle with the adjacent centered on an item position and of length c:

$$c = \sqrt[2]{2r^2(1-\cos\phi)}$$

where ϕ is the total angle divided by the number of items to be displayed and r is the distance from head to source. This improved design is illustrated in Fig. 4. The new approach is also robust against overshooting and it supports both minimal pointing gestures performed very close to the body and expressive gestures in which the arm is fully extended. As a secondary benefit, it could be used as a rapid scanning method when the hand is held close. When the arm is fully extended, angles are increased and the selection could be 'fine tuned'.

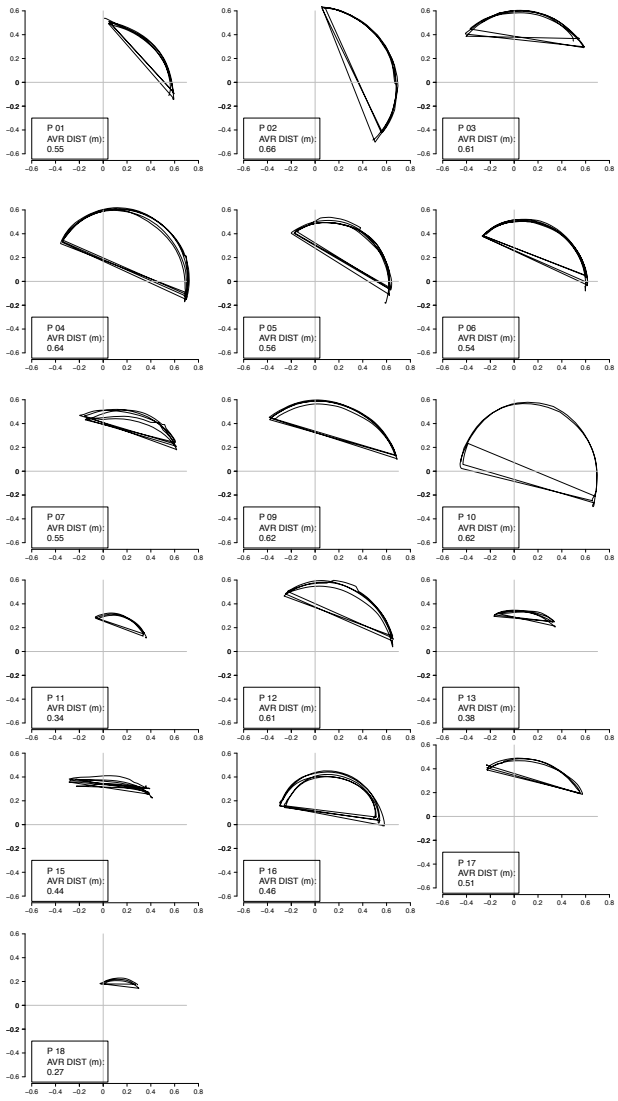

Figure 2: Individual angles circumscribed by right-handed participants (head at 0,0). Variations in preferred head to hand distance and start- and end- positions are apparent. On average, participants covered a range of 105.32 degrees and started at -15.39 degrees. The mean distance between hand and head was 0.55 meters.

[1] http://www.naturalpoint.com/optitrack
[2] http://processing.org
[3] http://puredata.info
[4] http://www.tu-berlin.de/?id=ssr

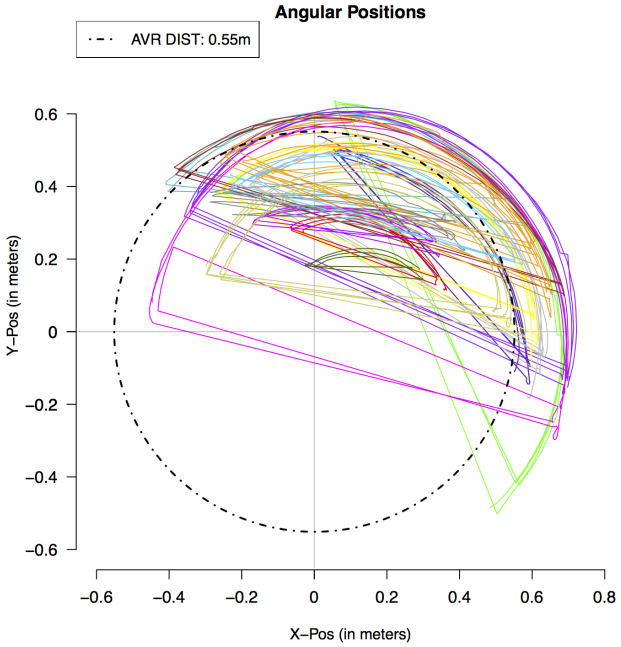

Figure 3: An overlay all angles circumscribed by participants. On average participants started at -15.39 degrees and covered a range of 105.32 degrees. The mean distance between hand and head (at 0,0) was 0.55 meters.

8. STUDY 2A: IMPACT OF PLAYBACK TYPE

The purpose of this study was to get a thorough understanding of how long on average it takes to select an item, how error prone this is, and how these two aspects are influenced by the playback type. The research question addressed in this study was:
RQ: What is the impact of the sound rendering (monophonic vs. binaural/spatial) on task completion time and error rate?

8.1 Experimental design

We compared two conditions in a within-subjects design:

- Cond. 1: 10 items, monophonic playback
- Cond. 2: 10 items, binaural playback

Participants completed both conditions in a counterbalanced order to prevent learning effects. Each condition consisted of 34 trials. At the beginning of a condition 10 musicons were randomly chosen from a total of 60 musicons and added to the list. The order of items was not changed during a condition but the target musicons were randomly picked from the current list. Dependent variables were *task completion time* and *error rate*. Independent variables were the *playback type of the sources pointed at (monophonic vs. binaural)*.

Before the experiment, all participants trained until they had a correct understanding of the procedure and the functionality of the equipment (usually 4 to 5 trials). Participants wore a glove to track their hands' position and register the pinch gesture. They also wore AKG K601 reference headphones with compensation filters applied, and equipped with optical markers to track their head's position and orientation. Between conditions, participants had breaks of 5 minutes. They completed the study in 30 minutes or less.

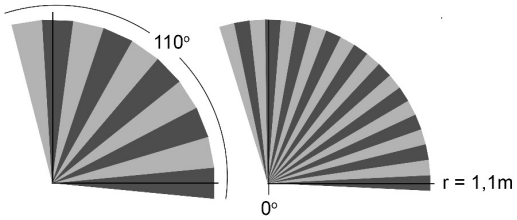

Figure 4: An illustration of the display layout used for 10-item conditions (left) and the 20-item condition (right). Items were arranged in a 110 degree arc starting at 11 o'clock. Participants could make a selection by pointing at an item, which was displayed at a distance of 1.1 meters.

8.2 Task and stimulus design

To address the research question participants had to find a song in a list of 10 songs, which were plaid to them either monophonically or binarually with spatial rendering. For the playback of songs we chose an approach suggested by [21] and created musicons–short characteristic samples taken from popular songs–to increase recognition rates and compensate for songs with quiet beginnings. Musicons were extracted from a total of 60 popular songs, taking the most characteristic samples with a length of six seconds, such as the first six seconds of Nirvana's *Smells Like Teen Spirit* or the refrain from Joan Jett & the Blackhearts' *I love Rock 'n' Roll*.

Participants could start a trial by pressing a button on a wireless presenter held in their left hand. Each trial began with a short *beep* followed by the target musicon played monophonically. A second *beep* signaled participants to begin the task. As illustrated in Fig. 1 (Left) participants could search the list by pointing their hand or finger at an item. Based on our previous findings, items were arranged in a 110 degree arc centred on the user's position and expanding from -30 degrees to 90 degrees (as shown in Fig. 4, top view). Pointing at an item started the playback of the musicon and stopped the previous. Participants could point at any item in any order, extend their arm fully or keep their hand close to their body. Once the target item was identified, participants could pinch their fingers (as illustrated in Fig. 1, Right) and mark the item. A short feedback sound played and the timer stopped.

8.3 Participants

16 right-handed participants with normal hearing from our institutions' database participated in the study. These 6 men and 10 women were between 25 and 68 years old (mean age 33 years). After the study they were compensated with a € 10 voucher.

8.4 Results

An independent-samples t-test was conducted to compare mean task completion times in the monophonic and binaural condition. For the analysis, we removed outliers with task completion times above 60 seconds and missing values. There was no significant difference in the scores for monophonic (M=5050 msec, SD=5576 msec) and binaural (M=5479 msec, SD=5481) playback, t(1054)=-1.161, p=.21. We conclude that the playback type had no statistical impact on task completion times.

We found a similar distribution in error rates. Overall, error rates were very low for all conditions. In condition monophonic participants made a total of 20 errors, 21 errors in the binaural condition.

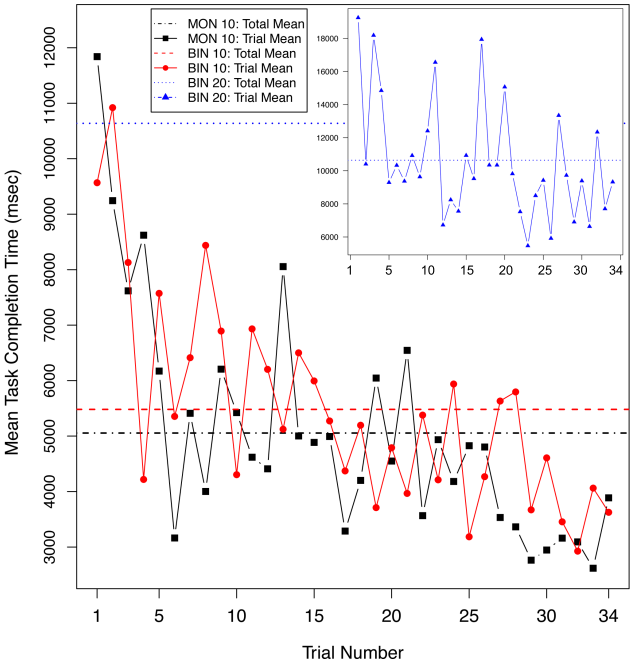

Figure 5: Mean task completion times (horizontal lines) for all conditions by trial number starting at trial 1. The decline reveals a learning effect equally strong for binaural (10) and monophonic (10) playback and not as pronounced for binaural (20), which is rendered in the upper right corner. (Note: data points are connected for better readability although measurement is not continuous).

9. STUDY 2B: IMPACT OF LIST LENGTH

In this follow-up study we were curious to learn about the impact of item numbers on task completion times and error rates. The research question addressed in this study is:
RQ: How are task completion time and error rate impacted by either 10 or 20 items in the list?

9.1 Experimental design

Study 2B had the same design, technical setup, number of participants and participants than study 2A, which is described in the previous section. The difference is that we compared a list of 10 items with a list of 20 items in a within-subjects design:

- Cond. 1: 10 items, binaural playback
- Cond. 2: 20 items, binaural playback

Dependent variables were *task completion time* and *error rate*. The independent variable was the *the number of items (10 vs. 20)* . Fig. 4 illustrates the different arrangement of items from a top view.

9.2 Results

An independent-samples t-test was conducted to compare mean task completion times in the 10 items and 20 items condition. For the analysis, we removed outliers with task completion times above 60 seconds and missing values. We found a significant difference in the scores for 10 items (M=5486 msec, SD=5485 msec) and 20 items (M=10668 msec, SD=12520), t(1038)=-8.582, p<.001. We conclude that the number of items has a statistical impact on task completion times. Participants were almost twice as fast in the 10 item condition than in the 20 item condition.

Again, error rates were low for the 10 item condition (19 errors). 60 errors were made in the 20 item condition showing a statistical difference from the 10-item condition with χ^2 (1, N=1040)=31.51, p<.001) but show only a small effect size (Cramér's V=.15).

9.3 Discussion

Summarizing these results, we found that in most cases it took participants less than 5.5 seconds to select an item from an unordered list of 10 items. As expected, task completion times increased when the list contained 20 items to almost twice as long as in the fastest condition (monophonic, 10). As plotted in Fig. 5, over time we see a tendency for decreased task completion times for 10-item conditions, though the effect was less pronounced for the 20-item condition. Knowing where an item is in the list may help locating it faster and hence lead to faster task completion times with increasing trial numbers. We also found that playing musicons from their position in space does not decrease task completion times. If a spatial memory effect exists, i.e. participants remember where musicons are located in the list and become faster at finding, time and the physical pointing gesture are likely to have a much stronger impact than the spatialized playback. When asked about their strategy, most participants explained they memorized the position of some well known or distinct musicons. Some also mentioned *regions* they associated with a musicon, like one participant described "I knew it was somewhere close to the end so I moved my hand there first and started searching nearby".
Error rates were surprisingly low given that participants had to hold their hand at an angle smaller than 11 degrees in the 10-item conditions and 5.5 degrees in the 20-item condition.
Although no statistical difference was found between the two 10-item conditions, participants tended to find the binaural condition easier to use.

10. STUDY 3: EVALUATION IN AN APPLIED CONTEXT

We built a simple music player to embody what we learned about the presentation and design of lists in auditory displays. We conducted a qualitative user study to look at how participants interacted with a list in this context. We also implemented some features accessed with interactions from our design space of lists, and tested how participants used and understood these features and the underlying concepts.

10.1 Design and features

To design a music player, we first identified typical tasks a user should be able to perform with it. We focused on: *selecting a song*, *playing/stoping a song*, and *adjusting the song's volume*. We used the general layout of the display from studies 2A and 2B but allowed users to chose from 60 alphabetically ordered songs. These were divided into three sections of each 20 songs and arranged evenly on a 110 degree arc at a distance of 1.1m (Fig. 4, Right).

10.1.1 Selecting an audio file

Sources pointed at played binaurally from their position for as long as the user's hand was in the respective angle segment (see Fig. 1, left). Touching thumb and index finger in a pinching gesture grabs the musicon, which stays at the hand's position for as long as fingers are pinched (see Fig. 1, right). Because it jumped from its original position (at a distance of 1.1 meters) to the user's hand's position its volume is now slightly increased.

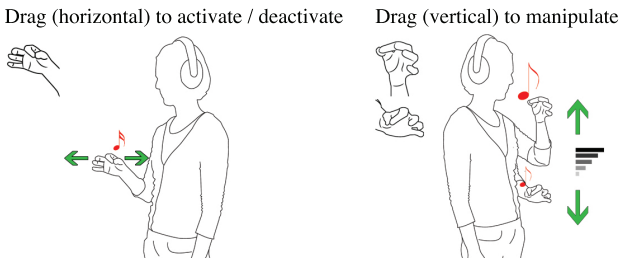

Figure 6: Gesture: Left: Drag item horizontally to activate or deactivate. Right: Drag item vertically to change the playback volume.

10.1.2 Playing and stopping an audio file

As illustrated in Fig. 6 (left), a musicon can be grabbed and pulled into a zone around the user, to play the song (in stereo). A song stops playing when it is pulled out of the play-zone.

10.1.3 Adjusting the volume

We implemented three different ways to change the volume. A user can: (1) grab a playing song and increase or decrease its volume by continuously moving the hand up or down and setting the volume upon release of the song (as illustrated in Fig. 6, right), (2) pinch and release fingers at different heights in the play-zone to stepwise increase or decrease the volume, (3) pull a musicon into the play-zone at a distinct height. Technically, we mapped absolute positions to absolute volume to help users associate regions of their body with loudness. A second reason we decided against a relative mapping was that without a reference point it is much easier to make harmful changes to the volume by accident.

10.1.4 Translation and rotation

We compared two versions: in the first, the display locked when the user's hand entered the play-zone, that is, translation remained fixed to the user but rotation was fixed to the world. In the second the display was not locked and both translation and rotation were user-fixed.

10.1.5 Scrolling

We implemented an approach similar to pagination to deal with a limit of around 20 items displayable in the arc. We divided a list of 60 alphabetically ordered songs into three sections. Pagination items were added at the start and end of each section. Users could page up or down by pinching this item. For example, when in section "F to P" the first item would page to songs in section "A to E" and the last would page to songs in section "R to Z".

10.1.6 Obtaining an overview

When a user paged to a new section, to help users gain an overview, we rapidly played all musicons consecutively for 700 msec from their spatial positions. The overview aborted when a user dropped the hand below the play-zone.

10.2 Experimental design and procedure

We evaluated the implementation in a qualitative user study using the thinking-aloud protocol. After we introduced participants to the player's general functionality they explored two versions–display locked and display unlocked (as described above in *Translation and rotation*)–for 10 minutes each. After the exploration period, the experimenter guided participants through each feature and participants shared their thoughts, questions, and recommendations.

10.2.1 Participants

We recruited 6 right-handed participants from our institution's database (3 male). They were between 24 and 34 years old (mean age 28 years) and received a small compensation for their time.

10.3 Results and discussion

When asked about their mental models of the list layout participants mentioned Apple's Cover Flow or being in the center of a wheel of fortune. In general they liked the idea of being surrounded by sound. The binaural rendering added to this perception and was appreciated by all participants. Although none of the participants had prior experience with binaurally rendered sound none had problems or was irritated by the idea of grabbing and manipulating invisible sound sources.

10.3.1 Selecting an audio file

All participants quickly learned how to select a song. Two participants would have preferred fewer songs per section to give them better control and more accuracy when pointing. Although participants were repeatedly reminded that they could hold their hand at any height and distance most kept their arm extended and pointed slightly upwards. A reason for this might be the rendering of sources at head height. Two Participants criticised the position and angular range of the display. One would have preferred range and position similar to his field of view while the other would have liked a smaller range starting at 12 o'clock.

10.3.2 Playing and stopping an audio file

Participants liked the size of the play-zone (approx. 20 cm from the center of the body, above waist level) and the grab and pull gestures. All participants noticed the increase in volume when they grabbed a musicon and it was played from the position of their hand. Some participants were first irritated when they pulled the musicon into the play-zone and a "different" song started playing. This happened when participants were not familiar with the song and hence could not match the musicon to the original song. When asked about feedback sounds for the grabbing gesture participants answered that the increase in volume and the change from musicon to song when entering the play-zone was sufficient. Participants used the stop gesture without problems. One participant combined the stop gesture in one smooth motion with lowering his hands to quiet the interface.

10.3.3 Adjusting the volume

Participants spent most of their time exploring different ways to adjust the volume and appreciated the direct change of volume when a song was grabbed and pulled up or down. When asked whether they would prefer a horizontal pulling gesture or a knob-turning gesture over the vertical adjustment, all participants favoured the vertical movement. Stated reasons for this were the familiarity with similar horizontal controls from operating systems like Windows or Mac OS. One participant mentioned the volume adjustment should set the master volume instead of each song's individual volume.

10.3.4 Translation and rotation

We asked participants to explore two versions of the implementation. All participants preferred the version in which the display is locked when the hand enters the play-zone and rotation is disabled. Reasons given for this were: (1) participants liked to use their body

instead of their head as reference system and found that it helped them to remember the spatial position of songs, (2) participants were keen to avoid conflicts between primary and secondary tasks in situations where orientational head movements would distract and make it difficult to target musicons, (3) participants felt more immersion in a "music space" when head rotations were compensated and physically pointing in a specific direction would always play the same song.

10.3.5 Scrolling

Participants could select songs from a total of 60 alphabetically ordered songs divided into three sections. Two participants initially had problems understanding the scrolling model. The main issue seemed to be the lack of analogies from similar interfaces. Two participants suggested a horizontal flicking gesture to "turn the wheel". Selecting a start or end item was the only event accompanied by a feedback sound. Most participants liked the "swooshing" sound and felt it illustrated the scrolling.

10.3.6 Obtaining an overview

We used rapid, successive playback of musicons to give an overview of songs in a list section. Most participants liked it as a feature but would have preferred to trigger it themselves instead of an automatic trigger when they paged to a new section. The option to interrupt the playback by lowering the hand below waist level was also appreciated. However, most participants said they would have preferred a simpler interrupt gesture like a pinch or flick gesture. All participants found a playback time of 700 msec to 1000 msec sufficient

11. CONCLUSION

We presented four user studies exploring a direct manipulation approach utilizing mid-air gestures to interact with spatialized lists in an auditory display. We learned that a list in the shape of a 110 degree arc angled towards the dominant hand is a comfortable and usable layout for most users. We showed that a selection takes less than 11 seconds and error rates are negligible when users locate an item in an unordered list of 20 items. As an example application we implemented a music player controlled with mid-air gestures. Users found the music player to be fun to interact with and had - in general - no problems selecting a song from a list of 60 alphabetically ordered songs. Our results suggest that mid-air gestures can be an efficient way to interact with lists in auditory displays such as playlist in simple music players. Although we used high precision tracking technology in our studies we believe that tracking technology, like proposed in [33] and applied more recently in [15, 5], show good tracking approaches beyond a fixed lab setup. We believe some of our findings are independent of the display's modality (auditory or visual) and could generally contribute to the design of gesture-based interactions using lists. Interesting application areas could be devices with limited I/O capabilities, three-dimensional, immersive environments like games, or interface alternatives for visually impaired users.

12. REFERENCES

[1] Ahrens, J., Geier, M., and Spors, S. The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *Audio Engineering Society Convention 124* (May 2008).

[2] Annett, J. On knowing how to do things: a theory of motor imagery. *Brain Res Cogn Brain Res 3*, 2 (March 1996), 65–69.

[3] Ashbrook, D., Baudisch, P., and White, S. Nenya: Subtle and eyes-free mobile input with a magnetically-tracked finger ring. In *Proceedings of SIGCHI*, CHI '11, ACM (New York, NY, USA, 2011), 2043–2046.

[4] Ashbrook, D. L. *Enabling mobile microinteractions*. PhD thesis, Atlanta, GA, USA, 2010. AAI3414437.

[5] Bailly, G., Müller, J., Rohs, M., Wigdor, D., and Kratz, S. Shoesense: a new perspective on gestural interaction and wearable applications. CHI '12, ACM (2012), 1239–1248.

[6] Baudel, T., and Beaudouin-Lafon, M. Charade: remote control of objects using free-hand gestures. *Commun. ACM 36*, 7 (July 1993), 28–35.

[7] Bernschütz, B., Stade, P., and Rühl, M. A spatial audio impulse response compilation captured at the wdr broadcast studios. In *27th Tonmeistertagung âĂŞ VDT International Convention* (Cologne, Germany, 2012).

[8] Brewster, S., and Raty, V.-P. Earcons as a method of providing navigational cues in a menu hierarchy. In *In Proceedings of BCS HCI'96*, Springer (1996), 169–183.

[9] Brewster, S. A., Wright, P. C., and Edwards, A. D. N. Parallel earcons: Reducing the length of audio messages. *IJHCS 43* (1995), 153–175.

[10] Cockburn, A., Quinn, P., Gutwin, C., Ramos, G., and Looser, J. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. *Int. J. Hum.-Comput. Stud. 69*, 6 (June 2011), 401–414.

[11] Crease, M., and Brewster, S. Making progress with sounds - the design and evaluation of an audio progress bar. In *British Computer Society* (1998), 167–177.

[12] Dicke, C., Deo, S., Billinghurst, M., Adams, N., and Lehikoinen, J. Experiments in mobile spatial audio-conferencing: key-based and gesture-based interaction. MobileHCI '08, ACM (2008), 91–100.

[13] Dicke, C., Wolf, K., and Tal, Y. Foogue: eyes-free interaction for smartphones. MobileHCI '10, ACM (2010), 455–458.

[14] Gamper, H., Dicke, C., Billinghurst, M., and Puolamäki, K. Sound sample detection and numerosity estimation using auditory display. *ACM Trans. Appl. Percept. 10*, 1 (mar 2013), 4:1–4:18.

[15] Gustafson, S., Bierwirth, D., and Baudisch, P. Imaginary interfaces: spatial interaction with empty hands and without visual feedback. UIST '10, ACM (2010), 3–12.

[16] Harrison, C., and Hudson, S. E. Abracadabra: Wireless, high-precision, and unpowered finger input for very small mobile devices. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, ACM (New York, NY, USA, 2009), 121–124.

[17] Kajastila, R., and Lokki, T. Eyes-free interaction with free-hand gestures and auditory menus. *Int. J. Hum.-Comput. Stud. 71*, 5 (may 2013), 627–640.

[18] Kildal, J., and Brewster, S. A. Non-visual overviews of complex data sets. In *CHI EA '06*, ACM (2006), 947–952.

[19] Magnusson, C., Rassmus-Gröhn, K., and Szymczak, D. Scanning angles for directional pointing. MobileHCI '10, ACM (2010), 399–400.

[20] Marentakis, G. N., and Brewster, S. A. A study on gestural interaction with a 3d audio display. In *Mobile HCI*, S. A. Brewster and M. D. Dunlop, Eds., vol. 3160 of *Lecture Notes in Computer Science*, Springer (2004), 180–191.

[21] McGee-Lennon, M., Wolters, M., McLachlan, R., Brewster, S., and Hall, C. Name that tune: musicons as reminders in the home. CHI '11, ACM (2011), 2803–2806.

[22] McGookin, D., Brewster, S., and Priego, P. Audio bubbles: Employing non-speech audio to support tourist wayfinding. HAID '09, Springer-Verlag (Berlin, Heidelberg, 2009), 41–50.

[23] Montero, C. S., Alexander, J., Marshall, M. T., and Subramanian, S. Would you do that?: understanding social acceptance of gestural interfaces. MobileHCI '10, ACM (2010), 275–278.

[24] Müller, J., Geier, M., Dicke, C., and Spors, S. The boomroom: Mid-air direct interaction with virtual sound sources. In *Proceedings of SIGCHI*, CHI '14, ACM (New York, NY, USA, 2014), 247–256.

[25] Oakley, I., and Park, J. Motion marking menus: An eyes-free approach to motion input for handheld devices. *Int. J. Hum.-Comput. Stud. 67*, 6 (June 2009), 515–532.

[26] Pielot, M., Kazakova, A., Hesselmann, T., Heuten, W., and Boll, S. Pocketmenu: non-visual menus for touch screen devices. MobileHCI '12, ACM (2012), 327–330.

[27] Pirhonen, A., Brewster, S., and Holguin, C. Gestural and audio metaphors as a means of control for mobile devices. CHI '02, ACM (2002), 291–298.

[28] Reeves, S., Benford, S., O'Malley, C., and Fraser, M. Designing the spectator experience. CHI '05, ACM (2005), 741–750.

[29] Rico, J., and Brewster, S. Usable gestures for mobile interfaces: evaluating social acceptability. CHI '10, ACM (2010), 887–896.

[30] Sawhney, N., and Schmandt, C. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput.-Hum. Interact. 7*, 3 (Sept. 2000), 353–383.

[31] Schmandt, C. Audio hallway: a virtual acoustic environment for browsing. UIST '98, ACM (1998), 163–170.

[32] Schwarz, J., Harrison, C., Hudson, S., and Mankoff, J. Cord input: An intuitive, high-accuracy, multi-degree-of-freedom input method for mobile devices. In *Proceedings of SIGCHI*, CHI '10, ACM (New York, NY, USA, 2010), 1657–1660.

[33] Segen, J., and Kumar, S. Gesture vr: vision-based 3d hand interace for spatial interaction. In *Proceedings of the sixth ACM international conference on Multimedia*, ACM (1998), 455–464.

[34] Sodnik, J., Dicke, C., Tomažič, S., and Billinghurst, M. A user study of auditory versus visual interfaces for use while driving. *IJHCS 66*, 5 (2008), 318–332.

[35] Stifelman, L. J., Arons, B., Schmandt, C., and Hulteen, E. A. Voicenotes: a speech interface for a hand-held voice notetaker. CHI '93, ACM (1993), 179–186.

[36] Strachan, S., Murray-Smith, R., and O'Modhrain, S. Bodyspace: inferring body pose for natural control of a music player. CHI EA '07, ACM (2007), 2001–2006.

[37] Vazquez-Alvarez, Y., Oakley, I., and Brewster, S. A. Auditory display design for exploration in mobile audio-augmented reality. *Personal Ubiquitous Comput. 16*, 8 (Dec. 2012), 987–999.

[38] Walker, B. N., Nance, A., and Lindsay, J. Spearcons: Speech-based earcons improve navigation performance in auditory menus. ICAD '06 (London, UK, 2006), 63–68.

[39] Wolf, K., Dicke, C., and Grasset, R. Touching the void: gestures for auditory interfaces. TEI '11, ACM (2011), 305–308.

[40] Yi, B., Cao, X., Fjeld, M., and Zhao, S. Exploring user motivations for eyes-free interaction on mobile devices. In *Proceedings of SIGCHI*, CHI '12, ACM (New York, NY, USA, 2012), 2789–2792.

[41] Zhao, S., Dragicevic, P., Chignell, M., Balakrishnan, R., and Baudisch, P. Earpod: eyes-free menu selection using touch input and reactive audio feedback. CHI '07, ACM (2007), 1395–1404.