The BoomRoom: Mid-air Direct Interaction with Virtual Sound Sources

Jörg Müller[1,2] **Matthias Geier**[3] **Christina Dicke**[2] **Sascha Spors**[3]
[1]Alexander von Humboldt Institute for Internet and Society, Berlin, Germany
[2]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[3]Universität Rostock, Rostock, Germany

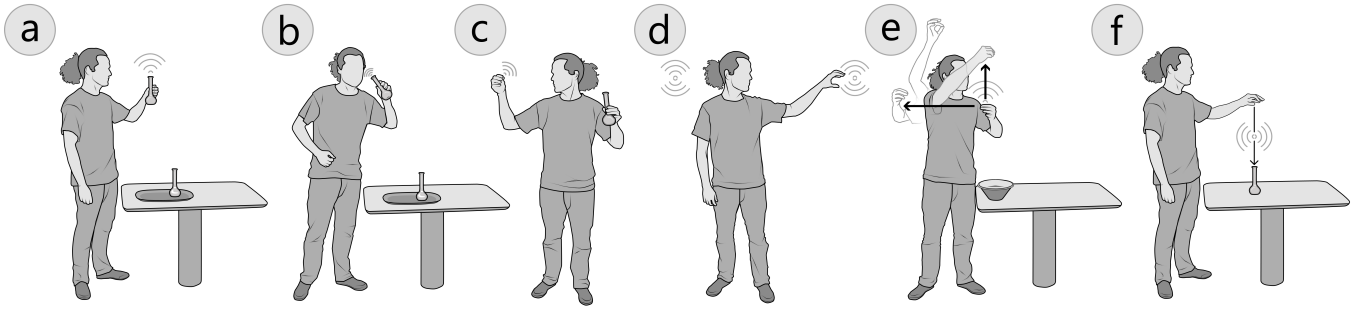

Figure 1. The BoomRoom allows to "touch", grab and manipulate sounds in mid-air. Further, real objects can seem to emit sound (a), even when being moved (b). Sounds can be picked up (c) and placed in mid-air (d). We use real world objects to augment the auditory feedback. For example, by using a bowl as filter object (e). Finally, sounds can be dropped into objects to be found more quickly (f). Sounds can be heard anywhere in the room, and appear to originate from the location of the virtual sound source regardless of the listeners position.

ABSTRACT

In this paper we present a system that allows to "touch", grab and manipulate sounds in mid-air. Further, arbitrary objects can seem to emit sound. We use spatial sound reproduction for sound rendering and computer vision for tracking. Using our approach, sounds can be heard from anywhere in the room and always appear to originate from the same (possibly moving) position, regardless of the listener's position. We demonstrate that direct "touch" interaction with sound is an interesting alternative to indirect interaction mediated through controllers or visual interfaces. We show that sound localization is surprisingly accurate (11.5 cm), even in the presence of distractors. We propose to leverage the ventriloquist effect to further increase localization accuracy. Finally, we demonstrate how affordances of real objects can create synergies of auditory and visual feedback. As an application of the system, we built a spatial music mixing room.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Mid-air; Spatial Sound Reproduction; Gestural Interaction.

CHI 2014, Apr 26 - May 01 2014, Toronto, ON, Canada ACM 978-1-4503-2473-1/14/04. http://dx.doi.org/10.1145/2556288.2557000

INTRODUCTION

As a hobby DJ, Marc has a BoomRoom installed in his living room. He invites his friend Laura, who is an amateur musician, for a jam session. Laura brings a few loops she has recorded with different instruments and uploads them into the system. Each instrument gets captured in a bottle on the table (Figure 1 a), so that Marc can pick up bottles and listen to them (b). He finds a sound that he likes and takes it out of the bottle (c). Meanwhile, Laura has taken a sound she particularly likes out of another bottle and hands it to Marc. Marc drops his sound in mid-air for later use and picks up the sound from Laura (d). He likes the sound, but explains to Laura that with a little bit of effect it could be even cooler. He walks over to his effect bowl, holds the sound over the bowl and stretches it with the other hand to distort it (e). They take a few of the other sounds and choose different variants, volumes, filters, etc. They place some sounds in mid-air, while they drop others into bottles (f) to create an interesting and engaging soundscape. They will continue to play with this soundscape at the party they are giving later that night.

In this paper we present the BoomRoom. The BoomRoom allows for direct manipulation of virtual sound sources hovering in mid-air. It also enables ordinary objects or body parts to appear to emit sounds. To accomplish this, BoomRoom uses a combination of spatial sound reproduction, in our case Wave Field Synthesis (WFS), and optical tracking. Loudspeakers and cameras can be at a distance from where the actual interaction takes place. We envision loudspeakers and cameras to be embedded into the walls and ceilings of arbitrary rooms. Further, we envision users to be completely uninstrumented, using the system as they are.

In our prototype we approximated this vision while simplifying our installation and increasing robustness. We created a small room (3 m diameter) where a circular array of 56 loudspeakers is hidden behind curtains. Further, we used a marker-based optical tracking system to simplify the computer vision part of gesture recognition, user tracking and object recognition and tracking.

In contrast to previous indirect interaction with audio, in this paper we propose to merge the location of the sound and of the interaction, enabling *auditory direct manipulation* of virtual sound sources hovering in mid-air.

We determine constraints and design issues of our system in three steps. First, we conducted two laboratory studies to determine important parameters of our system. We find the absolute pointing accuracy of our system to be 11.5 cm. Using distractors different from the stimulus, the accuracy degrades only insignificantly, but with similar distractors, accuracy degrades to up to 48 cm. To our knowledge this is the first direct pointing accuracy evaluation of a WFS system. Second, to showcase the capabilities and limitations of BoomRoom, we implemented a spatial music mixing application. Third, we provide learnings from theoretical considerations, our own experience, and a user-centered design process with invited novices and musicians.

This paper makes both a technical and a scientific contribution. On the technical side, we present the first system that allows to "touch", grab and manipulate sounds in mid-air. Further, arbitrary objects can seem to emit sound, even when moving. This is also the first WFS system that allows users to walk through a landscape of multiple, possibly moving, sounds in mid-air while always coping with the users' current head position.

On the scientific side, we present the first experiment using WFS that investigates how accurately users can "touch" a source. To our knowledge, previous experiments have only investigated how exactly users can point into the direction of a source. Finally, we present the first investigation of accuracy of WFS reproduction with distractors. These basic studies are necessary for a wide variety of interactions with sounds hovering in mid-air.

We took four main learnings from this project. First, we learned that direct "touch" interaction with sound is an interesting alternative to indirect interaction mediated by controllers or visual interfaces. It avoids a modality switch between auditory and visual modality. Further, it is very easy to learn by observation, and users describe it as natural and fun. Second, sound localization is surprisingly accurate (11.5 cm), even in the presence of distractors. However, the simultaneous presentation of very similar sounds should be avoided. Third, the localization cues of the visual and proprioceptive senses are stronger than the auditory cue. For sources close to the loudspeakers, the ventriloquist effect can create an unwanted bias towards the loudspeakers, however, the same effect helps to improve the impression that sounds are emerging from the users' hands or from physical objects. Fourth, spatial audio presents a limited bandwidth for feedback of gestural interaction. Therefore, affordances of real objects should be used to provide additional visual feedback for more complex interactions.

SCENARIOS

We believe that the ability to "touch" sound sources in mid-air and to make objects "speak" opens many new opportunities for human-computer interaction (HCI). As a simple example, the marble answering machine [11] could be taken to a new dimension. An ordinary bowl with marbles could be programmed to serve as an answering machine, making an occasional clicking sound by which the number of new messages is audible if a user is nearby. When a marble is taken out of the bowl, the marble itself could play the recorded message, while being carried through the room. If the user wants to delete the message, she could simply pull it out of the marble and drop it into the bin. He could even speak a reply into the marble that would be returned to the caller. If she wants to keep the message, she could simply drop the marble into another bowl. More generally, there would be no need for any devices in the room, like alarm clocks, bells, egg boilers, phones or computers, to have loudspeakers for themselves. Extending the vision of Audio Aura [16], unread emails could be a flock of birds that sit or fly somewhere in the room, with new mails flying in through the door and making a pass around the user. Urgent mails could occasionally fly over the user. By the chirp of the birds different senders could be recognized. If the user wants to read one of the messages, she could walk over to the bird, "touch" it, and the message would be read to her. By grabbing and manipulating the chirp, she could reply to or forward emails. As another example, smart rooms could finally become accessible for the blind. If a person comes into the room and wants to get an overview of the present objects, she could simply call the announce function and all objects would quickly announce themselves (keys, table, chair, etc.). Similarly, dropped objects on the floor or spilled liquids could make an appropriate sound to be detected by the user. Blind users could also simply attach their own sounds to objects by putting them into the objects, or leave messages for each other in mid-air.

SPATIAL AUDIO

The capabilities of the human auditory system to analyze acoustic scenes rely on the acoustic scattering of the outer ear including the upper torso, head and pinnae [2]. These acoustic properties are represented by so-called head-related transfer functions (HRTFs), which are dependent on distance and angle of incidence of a sound source. They are individual for each person. Humans can perceive sound coming from any direction; however, the localization accuracy depends on the spatial origin of the sound in relation to the position of the listener. The angular resolution is about 1–5 degrees of azimuth in front of the listener and up to 20 degrees for more peripheral and rear positions depending on the characteristics of the source and the presence of distractors [2, 24]. The localization accuracy in the median plane is much worse than in the horizontal plane.

Sound field synthesis (SFS) techniques aim at physically synthesizing a desired sound field within a defined listening area

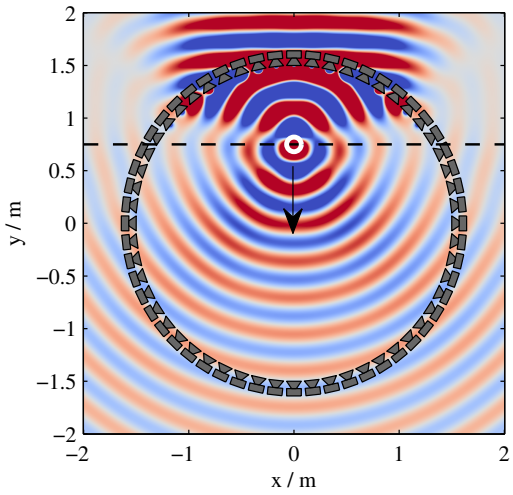

Figure 2. Synthesis of a focused source with WFS using a circular loudspeaker array of 56 loudspeakers. The array has a diameter of 3 m, the focused source emits a monochromatic sound field of 1 kHz and is located at (0, 0.75) m as indicated by the white ring. The sound impression is correct for any user in the lower area delimited by the dashed line.

using surrounding loudspeakers. Well known representatives are Wave Field Synthesis [1] and higher-order Ambisonics [5]. By synthesizing a sound field, the individual HRTFs of the listeners are preserved even when moving throughout the listening area. This is especially of interest in the presented work, since distance perception for nearby sources is related to individual spectral changes in the HRTFs [4].

In the following, we will focus on the foundations and practical limitations of WFS since this technique is used in the BoomRoom. The physical foundations of WFS are related to the Kirchhoff–Helmholtz integral from which WFS can be derived as stationary-phase approximation [25]. WFS allows for the synthesis of recorded or prescribed sound fields. The latter case is used commonly in *model-based rendering* where spatio-temporal models of virtual sources are used to derive the driving signals for the loudspeakers. Typical models are plane waves, point sources and focused sources. The former two constitute an acoustic source at an infinite respectively finite distance outside of the listening area. The latter one is the special case of an acoustic source located within the listening area. The BoomRoom uses only focused sources. Figure 2 illustrates the synthesis of a monochromatic focused source with a circular loudspeaker array in the same configuration as used in the BoomRoom. A sound field is synthesized that converges towards a focus point and diverges after the focus point at $(0, 0.75)$ m as a point source located at the focus. Hence, from a physical point of view the synthesized sound field is only correct for $y <= 0.75$ m within the loudspeaker array. Every listener located anywhere inside this area gets the impression of a sound emerging from the focus point.

The implementation of WFS faces some practical limitations. The theoretical background assumes a spatially continuous distribution of secondary sources. In practice, a finite number of loudspeakers is used, which constitutes a spatial sampling process. Typical loudspeaker spacings of 10 to 30 centimeters result in spatial sampling artifacts emerging for frequencies above 1 kHz. The results of psychoacoustic studies [20] and the considerable number of realized systems [6] prove that WFS allows for accurate spatial perception of sound scenes. The perceptual properties of focused sources have been investigated e.g. in [23, 22].

One problem that had to be tackled in the presented work is the limited listening area for focused sources. A listener who is located in the converging part of the sound field does not perceive the intended spatial impression [23]. While the sound itself is reproduced without major artifacts, it is located towards the loudspeakers. The propagation direction of a focused source can be controlled by sensible selection of active secondary sources [19]. Choosing the propagation direction towards the actual position of the listener, resolves the issue of the limited listening area. This holds also for multiple focused sources and for multiple listeners as long as no listener is located in the converging part of the sound field.

RELATED WORK

We review the state of the art regarding interaction with spatial audio on headphones and using loudspeakers.

Headphones

The vast majority of spatial audio work in HCI uses headphones. As an example, Brewster et al. [3] present two spatial audio interaction techniques to be used with headphones while walking. One is nodding into the direction of the sound source, while the second consists of gestural commands on a belt-mounted PDA. Audio Aura [16] augmented an office with non-spatial audio on headphones, such as sonifying emails or reminders. Strengths of using headphones are 1) mobility, 2) isolation from ambient noise, and 3) ability to render different sounds to different users. The major drawback is the necessity to wear headphones in the first place, which may be cumbersome and influence hearing, thereby separating the user from real-world sounds and other people.

Loudspeakers

The majority of research on WFS and related techniques concerns non-interactive spatial audio rendering. Typical applications are television and cinema, where they could provide the next generation of surround-sound which is not dependent on a sweet spot. In this section we discuss the few examples of gestural interactive WFS systems we could find.

Grainstick [13] is a gesture-controlled musical instrument using WFS. Two users stand in front of a linear loudspeaker array with optically tracked Wiimote controllers. The relative height of the two controllers controls virtual grains rendered as focused sources which move from one direction to the other in front of the WFS system.

The application of WFS in the context of visual Augmented Reality is discussed in [14]. The user is wearing video see-through AR glasses while standing inside a WFS system. The user can use a large paddle with a visual marker attached to the end to position a sound source. Seemingly, the sound source is permanently attached to the end of the paddle. In the same paper, also an exocentric perspective is presented,

where users look at the room as World-in-Miniature and can use a miniature paddle to position the sound source.

Springer et al. [21] present a system that combines WFS with a multi-viewer stereo display. Users stand in front of a large two-viewer stereo projection wearing a combination of shutter and polarization glasses. Behind the screen is a WFS system comprising three linear array segments. In one application, users hold an optically tracked controller to operate a billard cue on the screen. They can hit balls which bounce off cushions and other balls and thereby emit sounds. In another application, users hold a hand-held trackball. On the screen, a forest with a brook flowing from left to right is shown and audible. Users control a 3D cursor with the trackball, and when they press a button, a stone is dropped from the cursor into the brook.

Fohl et al. [7] present a gesture control interface for WFS. Users can point into the direction of a source and raise their hand to select it after a certain dwell time. When the hand is moved towards or away from the source, the source moves away from or towards the user. When the hand is moved sideways, the source rotates around the user. When the hand is dropped, the source is released. It is not stated in the paper whether focused sources are used, but since the head is not tracked, apparently users cannot walk around focused sources.

The main difference between these systems and our work is that users cannot "touch" the sources, but interact with them indirectly, by a) which of two controllers is higher [13] b) a paddle [14], c) a controller for pointing or a trackball [21] or d) the pointing direction of their hand [7]. Also different in our system is that users can walk freely around sources hovering in mid-air. The sources can even be moved around the user's own head, while a correct listening experience is maintained. This is apparently not possible in these previous systems. The ability to "touch" and move sources in mid-air is difficult to achieve without being able to walk around them. Finally, these systems provide mostly translation of sources, while we enable richer interaction, e.g., manipulation.

One critical extension of WFS that enables systems allowing to "touch" sounds is presented by Melchior et al. [15]. They do not present an interactive system, but rather a technique to select loudspeakers for focused sources based on the tracked listener position. This is a critical feature to enable users to walk around focused sources while continuously maintaining a correct listening impression. In the experiment, a physical (inactive) loudspeaker was placed in the center of a WFS system and users walked around it. They were asked whether they had the impression that the sound was coming from the loudspeaker (while it was actually rendered by the WFS system). We use the same approach to enable users to walk around sound sources in the BoomRoom. We extend the approach by 1) applying it to multiple sources simultaneously, so users can walk through an auditory landscape and by 2) enabling dynamically moving sources, so users can hold a source in their hand and move it around their head. We also 3) remove the physical prop, yielding the first system where users can walk with their head *through* a focused sound source. We discuss the consequences of these extensions in the paper.

In summary, while a few interactive WFS applications have been implemented, interaction is always *indirect*. In particular, we are not aware of any system where the users could "touch" the sound sources. This is partially due to the fact that in order to correctly render focused sources when the user is moving around, the approach of Melchior et al. is necessary. Further, in order to create sources anywhere in a room, a closed loudspeaker array is needed that encircles the entire room. Naturally, because "touching" sources has not been possible before, we present the first evaluations of "touch" accuracy in WFS. Finally, we present the first system that allows for more expressive interactions than mere indirect translation of sources.

THE IMPLEMENTATION OF BOOMROOM

The BoomRoom was realized in a room with a size of 4.3 m $\times$ 4.5 m. The room is equipped with absorber panels and curtains which reduce the reverberation time T_{60} to 0.5 seconds. In the room, a ring of 56 loudspeakers (Elac 301) is suspended from the ceiling. The ring has a diameter of 3 m which yields a loudspeaker spacing of about 17 cm. The height of the ring of loudspeakers can be varied; for the BoomRoom it was positioned at ear level of a standing person.

The loudspeaker driving signals are generated in real-time by a computer running the Debian GNU/Linux operating system. The model-based spatial audio reproduction was realized with the open-source software *SoundScape Renderer* (SSR) [8]. The SSR provides, among several other reproduction methods, a very efficient real-time implementation of WFS. The WFS algorithm is implemented using a delay line and a weighting factor for each source–loudspeaker pair and a static filter per source [6]. This so-called *pre-equalization filter* must be used to compensate the inherent low-pass characteristic of a loudspeaker array. It depends only on the layout of the loudspeaker array and is applied to each source signal before applying time delays and amplitude weighting factors. With a loudspeaker setup limited to the horizontal plane, the amplitude of the sound field cannot be synthesized correctly for the whole listening area. Therefore, a certain point inside the listening area is chosen as a reference point for the calculation of the amplitude. This reference point is typically located in the center of the loudspeaker array. For the BoomRoom, the SSR was extended with a feature called *reference offset*. This extension is publicly available in the latest release of the SSR.[1] The reference offset is bound to the tracked position of the listener's head, therefore the amplitude is always correctly calculated for the actual position of the listener. As mentioned above, for a given source only a subset of loudspeakers is used. This selection is also controlled by the reference offset and updated in real-time.

The interactive playback and looping of audio files and their routing to virtual sound sources in the SSR was realized with the visual programming language Pure Data (Pd). For the second experiment and the music mixing application (see

[1] `http://spatialaudio.net/ssr/`

below), several audio tracks have to be played back synchronously. This is done by loading a multi-channel audio file in Pd. At the end of the file, the playback is seamlessly started from the beginning. In addition to the instrumental loops, further audio files can be loaded for providing audio feedback in the music mixing application. These files can be started on demand and their position in the virtual room can be specified separately.

For the music mixing application two sound effects were implemented in Pd. One effect is a band-pass filter with resonance, where the cutoff frequency and the sharpness can be remote-controlled. The other effect is a simple distortion effect realized by wave-shaping the signal with a tangens hyperbolicus curve. The amount of distortion can be remote-controlled.

For optical tracking an OptiTrack system with 16 cameras suspended from the ceiling is used. We use pinch gloves to robustly detect pinches. Cap, gloves, and objects are equipped with reflective markers. The main system logic is implemented in Processing.The communication between applications was realized using TCP/IP sockets. For routing audio signals to the soundcard and between applications the JACK Audio Connection Kit was used.

EXPERIMENTAL EVALUATION 1: ACCURACY OF LOCALIZATION

The purpose of this study is to determine the accuracy with which users can locate virtual sound sources within our apparatus. This information is used to determine the radius within which a sound can be selected.

Procedure

We used the WFS apparatus described in the previous section. The head of the user was tracked with an optical marker. The selector was a Logitech Presenter with an optical marker attached. The participants selected using a button on the presenter.

Participants began a trial standing within the circle of loudspeakers by clicking the button. They heard a sound placed at a random position with at least 30 cm distance from the loudspeakers. Participants were able to move freely around the room within the loudspeaker array. Their task was to place the selector directly at the location where the sound appeared to be coming from and click the button. When they clicked the button, the sound was placed in a new random location, beginning the next trial. Users trained for 8 trials at the beginning of the experiment. The independent variable was the stimulus.

As stimuli, pulsed noise, speech, and guitar tones were used. All stimuli are available on the project website[2]. As dependent variables, the selection time from stimulus presentation to button press was measured and the distance between the selector and the sound source (projected to the horizontal plane) was continuously recorded over time. After each trial, the following stimulus was selected randomly. Users completed 3 stimuli $\times$ 15 repetitions = 45 trials. After all trials, a semi-structured interview regarding the user's strategy was performed.

[2] `http://joergmueller.info/boomroom/`

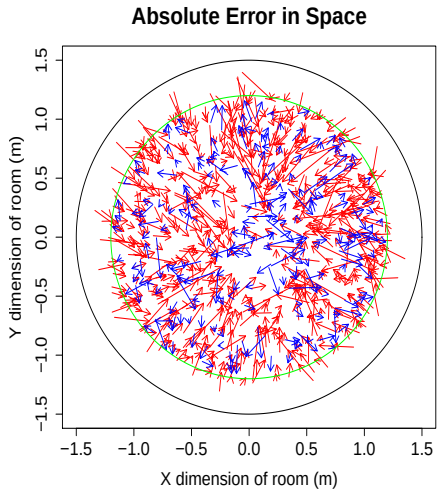

Figure 3. Error in absolute space. Arrows point from click to source. Red arrows indicate clicking towards the loudspeakers, while blue arrows indicate clicking towards the center of the room. As can be seen from the plot, there is a systematic bias for sources to be perceived as being closer to the loudspeakers (black circle, green circle denotes area where sources appeared).

We recruited 17 participants (6 male) to participate in the study. Participants were not associated with our laboratory and had no experience with spatial listening experiments or WFS before. They were aged between 25 and 68 (mean = 36). No participants reported any hearing impairments.

Results

The median accuracy (horizontal distance from click location to sound source) was 11.5 cm (mean = 14 cm). We did not find significant differences between stimuli (repeated-measures ANOVA, $F(2,715) = 1.39$, $p < .25$). We did find significant differences between participants, however (ANOVA, $F(16,748) = 21.9$, $p < .01$). Mean error of the most accurate participant was only 6.4 cm, while the least accurate participant had a mean error of 23.7 cm. Median selection time was 8 s.

When the stimulus appeared, participants needed about one second to localize it. Then they walked quickly towards it, reaching the vicinity of 30 cm after 3 s. They finally performed a fine search, where they improved their accuracy only slightly to 20 cm after 6 s.

The actual locations of the sound sources and the clicks are plotted in Figure 3. It can be seen that there is a systematic error for sources which are perceived to be closer to the curtain than they really are (t-test $t(764) = 13.7$, $p < .01$). The error from a user perspective is plotted in Figure 4. In this experiment, a slight tendency for overshooting (two-sided t-test $t(764) = -5.7$, $p < .01$) and bias to the right ($t(764) = 5.9$, $p < .01$) can be observed (all participants were using their right hand). Note also that the variance along the axis between head and source (SD = 14.6 cm) is significantly larger than

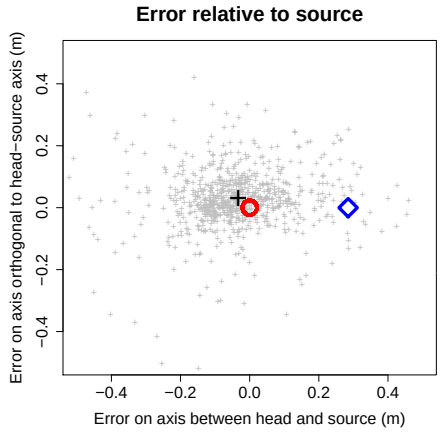

Figure 4. Error on an axis between source and head. Each trial is translated so that the source is at the center (red circle) and rotated so that the head is on the horizontal axis (mean head position blue diamond). Clicks are shown as grey + signs, the mean click location as black + sign. It can be seen that there is greater variance along the axis between head and source (distance estimation). Further, in our experiment participants tend to slightly overshoot the target and have a slight bias to their right.

the variance on the orthogonal axis (SD = 8.9 cm, $p < .01$, Bartlett test). It can also be seen that participants tend to have their head close to the source (median = 21.7 cm).

In interviews, participants described their strategy as first listening to the stimulus, rotating their head a bit, and then walking towards it. For the final approach, strategies differed. Some participants simply streched out their hand in the direction of the source when they believed to be close. Others described it as walking around the source. Others rotated their head to see on which ear the sound was louder. During the experiment a few users developed the strategy of walking through the source and determining when it jumped to the other side of their head. Many users tried to find the source with their head first. Then they either moved their head away and moved their hand to their previous head position, or they simply moved the hand very close to the head. Regarding the perceived location of the sound, many participants initially perceived the sound coming from the curtain. After a few trials, however, they reported to perceive the sound to originate in mid-air. Some participants also reported to perceive sources close to the curtains as coming from the curtains, and sources further towards the center of the room as originating in mid-air. Most participants expressed astonishment about their first ever experience of walking *through* a sound source. Some described the experience as a strange feeling in their head, as if the sound had entered their head. Others described it as the sound evading them, resulting in the perception of a moving sound source.

DISCUSSION

Localization of sound sources in mid-air is surprisingly accurate even for novice users (11.5 cm), and there are no significant differences between our stimuli. Determining the location of a sound from a distance or while walking around or even through sounds are different techniques, and we cannot distinguish their accuracy in this experiment. The two systematic misperceptions of a higher variance on the axis between head and source than the orthogonal axis, and the bias towards the loudspeakers, can be explained by psychoacoustics. Humans are much better in determining angles than distances. In our case, this effect is much less pronounced than in experiments where the listener is stationary [2]. Users employed active hearing, they translated and rotated their head. Even with this strategy, the effect is present. The perception of sources close to the curtains as coming from the curtains can be explained with visual dominance and the ventriloquist effect. If there is a plausible visual sound source (e.g., a curtain) close to an audible sound source, the visual perception may dominate the auditory perception, and the sound may be perceived as coming from the curtain.

EXPERIMENTAL EVALUATION 2: LOCALIZATION IN THE PRESENCE OF DISTRACTORS

The purpose of this study is to investigate the impact of varying numbers of distractors (both similar and dissimilar to the stimulus) on the accuracy of target acquisition.

Procedure

The same apparatus as in the previous experiment was used and the experiment was conducted immediately after the previous one with the same participants. Due to the application scenario, a prototypical implementation of a music mixer, musical instruments were chosen as stimuli for this study. Individual tracks from REM's song "It happened today", which are freely available under a Creative Commons license (CC BY-NC-SA 3.0), served as source material. An acoustic guitar was chosen as stimulus. As dissimilar distractors, different instruments (percussion, synth, mallets, bass, etc.) were used, and as similar distractors, different sounds from electric guitar, mandolin, and banjo. All sound files are available for download on the project website[2].

In the beginning of the experiment, participants could listen to the stimulus and all distractors separately. A trial began by listening to the stimulus in isolation. When the subject pressed a button, the stimulus changed location and 1, 3, 5 or 7 concurrent distractors (either similar or dissimilar) added at random locations (at least 20 cm distance from curtain) became audible. The task was to place the selector directly at the location where the stimulus appeared to be coming from and click the button. After the click, all distractors were muted, so that the subjects could estimate their own accuracy. 2 distractor categories $\times$ 4 different numbers of distractors $\times$ 3 repetitions = 24 trials were performed. After all trials, a semi-structured interview regarding the user's strategy was performed. As dependent variables, accuracy and time were measured.

Results

The average error for similar and dissimilar distractors by number of distractors is given in Figure 5. With a two-way ANOVA, we found significant main effects of kind of distractors ($F(1,78.7) = 63.08$, $p < .01$) and number of distractors ($F(3,78.7) = 7.89$, $p < .01$) on the performance time. We also found a significant interaction of kind of distractors and

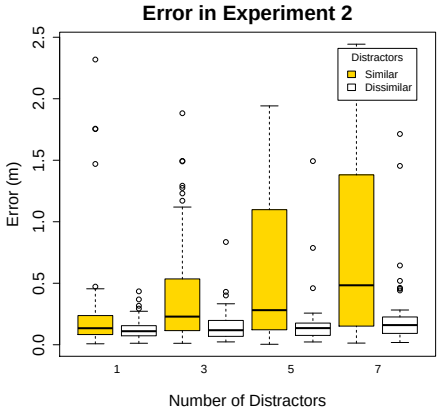

Figure 5. Error with similar and dissimilar distractors, for 1, 3, 5 and 7 distractors, respectively. For dissimilar distractors (white), the error is almost unaffected by the number of distractors. In contrast, with similar distractors (yellow) the system quickly becomes unusable when adding distractors.

number of distractors ($F(3,78.7) = 3.4$, $p < .05$). A Tukey's pairwise comparison revealed the significant differences between similar and dissimilar distractors as well as between 1-5, 1-7 and 3-7 distractors ($p < .05$). Median selection time was around 10 s and independent of the number of distractors. Bartlett's test found the variance in error to be higher for similar than for dissimilar distractors ($\chi^2(1) = 201.5$, $p < .01$). With dissimilar distractors, median errors for 1, 3, 5 and 7 distractors are 11 cm, 12 cm, 14 cm and 16 cm, respectively. Thus the system is still usable with distractors. With similar distractors however, the system quickly becomes unusable when adding distractors. Median errors for 1, 3, 5 and 7 distractors are 13 cm, 23 cm, 28 cm, and 48 cm, respectively. Further, median selection times rose sharply from 7.9 s for 1 distractor to 16.4 s for 7 distractors. Notably, with similar distractors, participants reported difficulties when multiple distractors were very close to each other or close to the stimulus. A two-way ANOVA found significant differences between participants ($F(16,374) = 4.37$, $p < .01$) and an interaction between participant and kind of distractors ($F(16,375) = 2.15$, $p < .01$). With similar distractors, the highest performing participant had a median error of 8.4 cm, while the least performing participant had a median error of 131.0 cm. With dissimilar distractors, this span was only 6.3 cm vs. 30.3 cm.

Discussion

While dissimilar distractors have little effect on performance, similar distractors make the system quickly unusable, both in terms of speed and accuracy. Participants did not report problems distinguishing the stimulus from distractors when presented separately. However, especially some participants were prone to confusing the stimulus with distractors when presented simultaneously. Concluding, while we do not see an issue presenting large numbers of dissimilar sounds concurrently, the concurrent presentation of similar sounds should be avoided if possible.

APPLICATIONS

In order to explore the capabilities of the BoomRoom, we implemented four different applications. We like to think about the BoomRoom to provide capabilities for consumption and creation. Regarding consumption, we implemented an application to augment the music listening experience. Instead of simply listening to a prefabricated stereo mix, the instruments and voices are distributed throughout the living room. E.g., violins may be situated close to the sofa, while flutes may hover above the table. Users can rearrange the spatial layout of the musical scenario at will. We have also implemented a lightsaber game where users can hold a small controller with two buttons. They can switch the lightsaber on and off, which then emits a lightsaber sound as if the saber would be about a meter long. Invisible enemies (which one can hear breathing) attack the players from all sides. The players have to defend themselves using the lightsaber. Third, we implemented an immersive audio drama experience where the voices occur from around the user. Here too, users can rearrange the scene at will.

In order to explore the creative capabilities of the BoomRoom, we implemented a spatial music mixing application. We were inspired by Ishii et al.'s examples of tangible computing [11], in particular the musicBottles [10]. With the musicBottles, sounds are confined within bottles placed on a dedicated table. Many users were seen to lift the bottles to their ears to hear whether the sound was literally coming from the bottle. However, this did not work, since the sound was coming from loudspeakers below the table.

We decided to take these concepts a step further. As with the musicBottles, musical instruments reside within bottles. In contrast to the musicBottles, the sound itself can be grasped and positioned somewhere else in the room. Sounds can also be dropped into bottles. In addition to bottles, there are a number of bowls in the room. The bowls can be programmed with arbitrary sound effects, and when a sound is held above a bowl, it can be altered. The sounds are explained to be elastic, so they can be held in place in one hand and stretched with the other hand in horizontal and/or vertical direction to be altered. We have implemented bowls for changing volume, selecting different variants of the instruments, applying filters and effects, making sounds play solo and muting them. For example, the volume of any sound can be changed simply by placing it above the volume bowl, holding it with one hand and stretching it vertically with the other hand.

EXPERIENCES WITH MID-AIR AUDITORY DIRECT MANIPULATION

We invited a dozen users to explore interaction with the music mixing application. Users came from different backgrounds (from no musical experience over audio experts to professional musicians). In this section, we provide learnings from theoretical considerations, from our own experiences and from this user-centered design process.

We present our results in form of a design space of primitive interactions with sounds hovering in mid-air. We identified five primitive interactions: Finding, selecting, grabbing, ma-

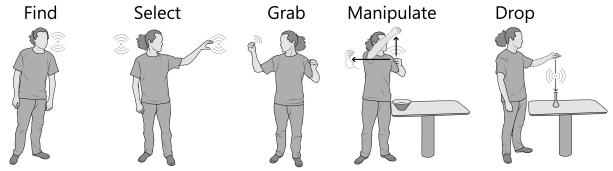

Figure 6. Design space of primitive interactions with sounds hovering in mid-air.

nipulating and dropping sounds. We argue that the combination of these interaction primitives enables direct manipulation of sound sources in mid-air, inspired by the concept of direct manipulation for graphical user interfaces. Direct manipulation interfaces [18] are charactized by 1) Continuous representation of the object of interest; 2) Physical actions or labeled button presses instead of complex syntax; 3) Rapid incremental reversible operations whose impact on the object of interest is immediately visible. They are argued to be beneficial in particular because of the closeness of the user's mental model and the physical requirements of the system, and because of the user's feeling of interacting with the objects themselves rather than via a tool [9]. Such interfaces are argued to provide benefits like ease of learning by imitation, the immediate feedback whether the actions are furthering one's goals, and the ability to simply change the direction of ones activity otherwise [18]. xperienced user. We discuss each of the interaction primitives in turn.

Finding

Finding involves determining the location of a specific sound, possibly within a complex scene. We observed that users rotate their head a lot and walk around in the room. In order to support these strategies, especially moving around sources, the extension of the method of Melchior [15], which was used in our implementation, is necessary.

One particularly interesting aspect of finding sounds relates to the first principle of direct manipulation, continuous representation of the object of interest. While objects in visual interfaces are usually visible continuously, audio is often not continuous. In silent moments, it may be difficult or impossible to either find a source or hear the effects when it is manipulated. We propose three solutions to this problem: (1) Modifying the source signal, (2) adding an announce function and (3) attaching the source to visual objects or body parts.

First, in our music mixing application, we simply removed most moments of silence in the source signal. The central downside of this approach is the modification of the audio scene itself, which may be inappropriate in the case of music or speech.

Second, we implemented an announce function. Sounds announce themselves via auditory icons (musicons in our case) or speech labels when the user lifts a hand above the head without pinching. In our case, all sounds in the room announce themselves. For a larger number of sounds, it would be beneficial if the user could select a region for items to annouce themselves via a torch metaphor. A torch could highlight sounds either in a cone (direction) or circle (position). We use overlapping announcement of sounds within 200 ms.

Third, we assigned sounds to objects or body parts (the hands), such that they can simply be found via the visual or proprioceptive senses. The user needs to remember which source is attached to which object, but then these sounds can be found efficiently using the visual or proprioceptive senses regardless whether they are currently audible.

Selecting

Selecting involves the determination of one sound for further interaction. In our case, selecting is performed by positioning one hand in an area around the sound. When the selection area is entered or left, a feedback sound is played. Important choices are the size of the selection area and feedback sounds. There is a general tradeoff between ease of selection and inadvertent selections, either when multiple sounds are close or e.g., while walking through the room. We observed that users often walk in the direction of the source and then sweep their hand in front of them until they hear the feedback that the source is selected. We currently use a horizontal radius of 20 cm for source size, which is well above the average localization accuracy of 11.5 cm and works well. In cases where users were unable to select a source, their head was often very close to the source, making it difficult to select. Equally important is the vertical size of the selection area. It can be very annoying to get a large number of entered/left feedbacks when one is walking around in the room. Therefore, it should be taken care that sounds are not selected when hands are not raised. We observed that when a sound is not attached to an object, users tend to lift their hands to the height of the loudspeaker array for selection. The vertical localization accuracy is enough for people to experience the vertical position of the sounds at the height of the loudspeakers. For sounds hovering in mid-air, we currently define the vertical selection area as starting 10 cm below the loudspeaker array.

For sounds that are attached to objects such as bottles, we leverage the ventriloquist effect for selection. The ventriloquist effect describes that if there is a discrepancy between auditory and visual localization cues, the perception is biased towards the visual cue [12]. The effect works well for angular deviations between auditory and visual cues of up to 20°–30° even if there is no apparent causal relationship between visual and auditory events [12]. It should be noted that in the literature only perceived differences in azimuthal angles are investigated. We are not aware of any experiments which investigate localization with regard to different elevation angles and different distances. One can assume, however, that the bias towards the visual cue is even greater in elevation and distance because auditory localization on its own is much less accurate in these cases [2].

In our experience, users have the perception that the sound is coming directly from the opening of the bottles, and that this experience is quite robust against vertical angular deviations between bottle and loudspeakers. For steep angles, as when standing close to the bottles, however, they have the feeling that the sound is hovering above the bottle. For sounds inside objects, we define a selection area that starts immediately above the object.

Grabbing

For grabbing, we use the pinch gesture because it is robust to recognize using computer vision and easy to understand. When a user pinches while a sound is selected, that sound is grabbed and can be moved. We currently give a general feedback that a sound is grabbed. However, users also need to verify which source they have grabbed. The behaviour we have observed most often was to move the sound close to the ear and back and forth, or to move it around the head. Using this technique, users were able to verify quickly which source they had grabbed, so that we see no need for additional feedback. Users also reported that the spatial impression from the audio was strongest when they had grabbed a sound. This can be explained with the proprioceptive ventriloquist effect. This effect explains that if there is a discrepancy between auditory and proprioceptive cues, the perception is biased towards the proprioceptive cues [17].

Manipulating

For manipulating sound, our first approach was to use mid-air gestures. In the first iteration of the volume adjustment, sounds could be grabbed in mid-air and then be moved up and down to raise and lower the volume. This approach, however, clashed with the proprioceptive ventriloquist effect. When the sound was grabbed and the hand lowered, there was a strong expectation that the sound should move vertically, too. Because the angular difference between loudspeakers and hand was so large, it became audible that the sound was still coming from the same location, breaking the illusion that the sound was held in hand. Further, we had problems with the limited feedback bandwidth of spatial audio compared to visual interfaces, making it difficult to communicate in which manipulation mode the user currently was. Our second approach involved physical objects, like a pepper caster, that could be held in hand and be "put inside the sound". This however strongly reduced the manual dexterity for gestural manipulations, in particular it was difficult to use pinch as a delimiter when holding an object. In our current approach, we use bowls, which have the affordance that sounds can be "put into" and "held over" them. The hands are now free for gestures and the bowls provide a visual feedback for the zone where each action can be performed. In order to maintain consistency with the proprioceptive ventriloquist effect, the sound always needs to be held in one hand above the bowl. The other hand can then define two parameters by moving horizontally and vertically.

We support three different manipulation styles: relative, absolute, and discrete manipulation. Relative manipulation is used for parameters that users usually want to manipulate relative to their current value, such as volume. In our initial implementation volume was defined by the relative height of the two hands. When both hands were at the same height, there was no change in volume. We quickly noticed that users had difficulty understanding this concept. Instead, most users grabbed the sound with one hand, pinched with the other hand at an arbitrary location, and moved the second hand up and down in the expectation for the volume to change accordingly. We subsequently implemented this behavior. For other parameters, it is important that they can be easily set to zero regardless of the current value, such as filter or effect. For these parameters, we use an absolute selection style, where both hands close to each other set the value to zero, and the distance in horizontal direction sets one parameter and the distance in vertical direction the other. Note that it is difficult to cross both hands, therefore we use the horizontal axis for parameters that have only positive values (such as filter steepness or effect strength) and the vertical direction for parameters that have both positive and negative values (such aus frequency relative to 440 Hz). For discrete parameters, such as variant, we quickly noticed that it was not always audible when the value/variant had changed. We therefore added discrete feedback for these events.

Dropping

Dropping simply involves releasing the pinch when a sound is grabbed. Sounds can be dropped in mid-air, into bottles or bowls. Dropping sounds into bowls allows to apply an effect to multiple sounds simultaneously (useful, e.g., for solo). A different feedback is given when sounds are dropped in mid-air or into objects, to give users the chance to verify that they have successfully dropped a sound into an object. In our first iteration, sounds were dropped into objects when they were close to the object and not grabbed. This lead to the phenomenon of inadvertently "collecting" all sounds along the way when one carried an object across the room. In our current implementation, sounds are only dropped into objects if they are explicitly released above them.

LIMITATIONS

The experiments were conducted with a single user at a time. The tracked position of the user's head was used as reference point for calculating the WFS driving signals. With a few limitations, the BoomRoom is also multiuser capable. In this case, the reference point is chosen between users [15]. Therefore, a focused source located directly between users cannot be rendered correctly for all users as some of them would be outside the "allowed" area (see Figure 2). Nevertheless, all other positions will work for all users. Furthermore, when one user is standing still while others are moving, the sound perception changes also for the stationary user because the reference point is changing. Finally, like all practical sound field reproduction systems, the BoomRoom suffers from more or less audible artifacts caused by spatial aliasing [20].

These are all physical limitations and not shortcomings of the current implementation. However, future research based on sound perception may lead to methods that allow to elude these physical limitations.

CONCLUSION

We presented a system that allows users to "touch", grab and manipulate sounds in mid-air. We took four main learnings from this project. We learned that direct "touch" interaction with sound is an interesting alternative to indirect interaction mediated by controllers or visual interfaces. Sound localization is surprisingly accurate (11.5 cm), even in the presence of distractors. The ventriloquist effect can be leveraged by assigning sounds to real objects or holding them in the

hands. Finally, affordances of real objects should be used to enrich the limited feedback bandwidth of spatial audio for more complex interactions. We believe that mid-air auditory direct manipulation has significant potential beyond what we explored in this paper.

ACKNOWLEDGEMENTS

We particularly want to thank Sean Gustafson and Patrick Baudisch for extensive discussions and support on this project. This work was supported by the ICT Labs of the European Institute of Innovation and Technology.

REFERENCES

1. Berkhout, A. A holographic approach to acoustic control. *Journal of the Audio Engineering Society 36*, 12 (1988), 977–995.
2. Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised ed. MIT Press, 1996.
3. Brewster, S., Lumsden, J., Bell, M., Hall, M., and Tasker, S. Multimodal 'eyes-free' interaction techniques for wearable devices. In *SIGCHI Conference on Human Factors in Computing Systems* (2003).
4. Brungart, D. S., and Rabinowitz, W. M. Auditory localization of nearby sources. Head-related transfer functions. *Journal of the Acoustical Socienty of America 106*, 3 (1999), 1465–1479.
5. Daniel, J. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format. In *23rd International Conference of the Audio Engineering Society* (2003).
6. de Vries, D. *Wave Field Synthesis*. AES Monograph. Audio Engineering Society, 2009.
7. Fohl, W., and Nogalski, M. A gesture control interface for a Wave Field Synthesis system. In *International Conference on New Interfaces for Musical Expression* (2013).
8. Geier, M., and Spors, S. Spatial audio reproduction with the SoundScape Renderer. In *27th Tonmeistertagung – VDT International Convention* (2012).
9. Hutchins, E. L., Hollan, J. D., and Norman, D. A. Direct manipulation interfaces. *Human–Computer Interaction 1*, 4 (1985), 311–338.
10. Ishii, H., Mazalek, A., and Lee, J. Bottles as a minimal interface to access digital information. In *SIGCHI Conference on Human Factors in Computing Systems* (2001).
11. Ishii, H., and Ullmer, B. Tangible bits: towards seamless interfaces between people, bits and atoms. In *SIGCHI Conference on Human Factors in Computing Systems* (1997).
12. Jackson, C. V. Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology 5*, 2 (1953), 52–65.
13. Leslie, G., Zamborlin, B., Jodlowski, P., and Schnell, N. Grainstick: A collaborative, interactive sound installation. In *International Computer Music Conference* (2010).
14. Melchior, F., Laubach, T., and de Vries, D. Authoring and user interaction for the production of Wave Field Synthesis content in an augmented reality system. In *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality* (2005).
15. Melchior, F., Sladeczek, C., de Vries, D., and Fröhlich, B. User-dependent optimization of Wave Field Synthesis reproduction for directive sound fields. In *124th Convention of the Audio Engineering Society* (2008).
16. Mynatt, E. D., Back, M., Want, R., Baer, M., and Ellis, J. B. Designing Audio Aura. In *SIGCHI Conference on Human Factors in Computing Systems* (1998).
17. Pick, H. L., Warren, D. H., and Hay, J. C. Sensory conflict in judgments of spatial direction. *Perception & Psychophysics 6*, 4 (1969), 203–205.
18. Shneiderman, B. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology 1*, 3 (1982), 237–256.
19. Spors, S. Extension of an analytic secondary source selection criterion for Wave Field Synthesis. In *123rd Convention of the Audio Engineering Society* (2007).
20. Spors, S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F. Spatial sound with loudspeakers and its perception: A review of the current state. *IEEE Proceedings 101*, 9 (2013), 1920–1938.
21. Springer, J. P., Sladeczek, C., Scheffler, M., Hochstrate, J., Melchior, F., and Frohlich, B. Combining Wave Field Synthesis and multi-viewer stereo displays. In *IEEE Virtual Reality Conference* (2006).
22. Völk, F., Mühlbauer, U., and Fastl, H. Minimum audible distance (MAD) by the example of Wave Field Synthesis. In *German Annual Conference on Acoustics (DAGA)* (2012).
23. Wierstorf, H., Raake, A., Geier, M., and Spors, S. Perception of focused sources in Wave Field Synthesis. *Journal of the Audio Engineering Society 61*, 1/2 (2013), 5–16.
24. Yost, W. A., Dye, R. H., and Sheft, S. A simulated cocktail party with up to three sound sources. *Perception & Psychophysics 58* (1996), 1026–1036.
25. Zotter, F., and Spors, S. Is sound field control determined at all frequencies? How is it related to numerical acoustics? In *52nd Conference of the Audio Engineering Society* (2013).