Human Interaction in 3D Manipulations

Can sonification improve the performance of the interaction?

Viking Edström - vedstrom@kth.se
Fredrik Hallberg - fhallb@kth.se

Stockholm 2014-04-29

CSC - TMH
Kungliga Tekniska Högskolan
Abstract

In this report the effects of using sonification when performing movements in 3D space are explored. User studies were performed where participants had to repeatedly move their hand toward a target. Three different sonification modes were tested where the fundamental frequency, sound level and sound rate were varied respectively depending on the distance to the target. The results show that there is no statistically significant performance increase for any sonification mode. There is however an indication that sonification increases the interaction speed for some users. The mode which provided the greatest average performance increase was when the sound level was varied. This mode gave a 7% average speed increase over the silent control mode. However, the sound level mode has some significant drawbacks, especially the very high base volume requirement, which might not make it the best suited sonification mode for all applications. In the general case we instead recommend using the sonification mode that varies the sound rate, which gave a slightly lower performance gain but can be played at a lower volume due to its binary nature.
Contents

1 Introduction 3

2 Problem formulation 3

3 Background 3

4 Method 4
  4.1 Hardware ................................................................. 4
  4.2 Software ................................................................. 5
  4.3 Sonification modes ....................................................... 5
    4.3.1 Frequency ............................................................ 5
    4.3.2 Volume ............................................................... 5
    4.3.3 Rate ................................................................. 6
  4.4 Participants ............................................................. 6
  4.5 Setup ................................................................. 7
  4.6 Testing procedure ..................................................... 7

5 Results 8
  5.1 Data exclusion ........................................................ 8
  5.2 Data ................................................................. 8
    5.2.1 Normal distribution plot ........................................... 11
  5.3 Observations ........................................................... 11
  5.4 Discussions .......................................................... 12
  5.5 Error sources ........................................................ 13

6 Conclusion 13

References 15
1 Introduction

In 1954 Paul Fitts formulated a law stating that the time required for a human to point to a target is a function of both the distance to the target and the size of the target. The law predicts that a small target will take longer time to reach than a big target at the same distance [1, 2]. Fitts law was formulated to mainly consider movement in a one dimensional space but several studies have shown that similar relationships occur in 2D and 3D [3, 4].

Fitts law does not explicitly take into account what kind of feedback the user gets, it only actively predicts the performance. The most common, and often easiest, way to relay information to a user is via vision, e.g. a pointer on a screen. However the use of other human senses to enhance performance and experience can and has been used for some time, although in a smaller scale and in other scenarios. One example is haptic feedback that is becoming increasingly common today when touch screens are involved. One use of this technology is to give the user a notification, via a slight vibration, that a touch has been registered.

Audio feedback is another interesting way of using one of the other basic human senses than vision to improve interaction. This type of feedback through the auditory system is called sonification. Only sounds which do not use voice or other natural languages qualify as sonification [5]. An example of auditory information which is not voice but still uses natural language is Morse code.

Combining visual feedback with feedback targeting other senses might give performance improvements and might provide a more pleasant user experience. Multimodal feedback might even enable completely new types of interactions in areas where one sense is occupied or limited.

2 Problem formulation

The computer mouse is by far the most popular way of controlling computers due to its precision and flexibility. The mouse has been around for some 50 years and works well for controlling a pointer on a 2-dimensional screen [8]. However as the mouse is limited to moving in only 2 axis and with 3D displays on the rise, other more modern input devices are being considered as replacements. One example is hands-off controls in 3D space. This method has been widely used in sci-fi movies as futuristic technology but has lately become common in living rooms as controllers for video games and other applications. The issue with this type of input is that it raises concerns about maintaining high precision while keeping interaction speed high. If the interaction is slower and more cumbersome than using a mouse for input it will probably never get adopted by a large user group and this might hinder the development. The question this report will try to answer is therefore: Can sonification improve the performance of human interaction when operating in 3D-space and if so, will different modes of sonification provide different levels of improvement?

3 Background

Some sonification already exists in modern interfaces, a good example is the sound made when reversing towards an object in a modern car, a sound that
consists of many beeps which rate increase as the object gets closer and finally becomes a continuous tone when the object is dangerously close. Another example is the click sound in older versions of Windows when the user presses a button. As users grew more and more accustomed to the interaction, and the system could respond faster, this sound was deprecated because it did not confer any additional information. The users already knew they had clicked without a sound indication. A counterexample is the shutter-sound in digital cameras (and now mobile phones), which imitates a traditional camera. Most people simply do not realize that they have taken a photo without this sound. Despite examples from both older and more modern applications, sonification is still a relatively unexplored field. [6]

The most basic type of sonification is rigid, event-based sound cues, such as a phone ringing. These often serve the purpose of getting the user’s attention when it might be elsewhere. This usage is not very interactive. Another type of sonification is using sound to represent a sequence of related data. Since sound can be modulated in many different ways, even highly dimensional data can be effectively represented. The human ear can simultaneously differentiate between many different characteristics of a sound. This is still not very interactive, since it is typically computed on a set or stream of data over which the user is not fully in control. This has applications in e.g. EEG (electroencephalography) analysis, stock market monitoring and other medical or financial data streams. With the right parameters the resulting sound sequences can even be considered music, and by varying the input different pieces can be generated [7]. When longer data series are converted to a sound track and played back to the user without any interaction, it is often called audification. [6]

The most interactive type of sonification is real-time sonification, which links one or more data quantities the user can directly manipulate to certain sound characteristics. This type of sound is used to highlight a change in quantity over time to the user such as score, distance or amount. Real-time sonification is usually a continuous stream of feedback, which is always active even if the state has not changed, like the aforementioned parking sensor. The scale from rigid to interactive sonification is very fuzzy and a sound can have characteristics from more than one of these types. [6]

4 Method

The problem formulation was researched by performing user studies. A software suite was created for testing performance of users doing movement in 3D space.

4.1 Hardware

The motion-tracking device used in this study was a Microsoft Kinect[1]. The Kinect uses two regular cameras and an IR-sensor to produce a 3D image of the room and provide body-tracking. While the Kinect in most cases had very good tracking performance, it did have some quirks. One was the existence of twitching, where the sensor quickly switches between two positions for the hand which made it almost impossible to indicate a single position accurately. Another issue was the inexactness and increased twitching which occurred when

trying to track the right hand while it was on the left side of the body or far away from the centre of mass.

Alongside, a projector was used to give the participant a good visual image of the software. To generate the required sound at a volume loud and clear enough a powerful speaker of the brand Harman Kardon was used.

4.2 Software

The software was created in Unity\(^2\), a game engine providing a 3D environment assuring a quicker progress in the software development phase. Integrating this to use the data from the Kinect was straight-forward since all involved interfaces were easy to use. Initial testing while programming concluded that the Kinect had poor performance (a lot of twitching) when the user’s right hand was on the left side of the body, the target placement ranges were therefore limited to the right side of the body.

On the screen the users see is a small white sphere representing their hand (the "hand") and a cube (the staging area) which is used as a common start position for all runs. The scene is set so that the user views his hand from behind in a slightly downward tilted angle. When moving the hand to the staging area, a red ball appears on the screen (the target). See figure 1 for a screen shot of the environment.

This target ball has a new random position - but does not change size - for every test run. The sonification mode used for each run is displayed as a text at the top of the screen. After holding the hand in the staging area for two seconds (indicated by a progress bar), the timer starts and the users should move their hand to the target as quickly as possible. The timer stops when the hand reaches the target and holds still within the ball for one continuous second.

4.3 Sonification modes

This section describes the different modes of sonification used in the testing. The modes were chosen because they represent three different ways a sound can change, change in frequency, change in volume and in the pattern of the sound (rate). All three modes operated on a base tone of 440 Hz (Concert A). This tone can be pitched up and down while staying within in hearing range.

4.3.1 Frequency

This sonification mode changes the frequency of the sound resulting in a higher pitch when approaching the target. When the hand is within the boundaries of the target the sound goes silent to indicate that the user has reached the target and should hold still. An exponentially growing function gave small fluctuations in distance a greater resulting difference in frequency closer to the target.

4.3.2 Volume

This sonification mode changes the volume of the sound, resulting in a higher volume when approaching the target. On reaching the target the sound was

\(^2\)http://www.unity3d.com
4.3.3 Rate

This sonification mode utilizes an increasing rate of isophase beeps as the user gets closer to the target. At the maximum distance (approximately the greatest distance the target can be placed from the staging area) the rate is about 3 beeps per second, increasing to 10 beeps per second right before the user hits the target at which point the sound becomes a constant tone. During development of the test software, the tone used for the beeps was initially a 440 Hz tone, but pilot tests indicated a more natural feel to the feedback if a higher pitch was used (660 Hz).

4.4 Participants

The people who were chosen to participate in the experiment were mostly friends of the authors. This meant a skew towards males with a high degree of computer skill. In total 9 men and 3 women took part in the study. Ages were quite tightly grouped in the 20-30 segment, with some outliers. A potential problem that could have affected the results were the handedness of the test participant since the box in which the target could be placed was not symmetrical around the staging area, but quite heavily skewed towards the right. Therefore all participants were required to be right-handed.
4.5 Setup

The user test was set up in a small room with the projector pointing towards a wall with the Kinect sitting right below it. The distance from the Kinect to the user was determined by the staging area in the 3D world. The users were told to stand so that their hand was in the staging area when it was casually hanging at their side. This happened approximately 275 cm away from the Kinect for a person of average height. See figure 2 for reference.

4.6 Testing procedure

To eliminate bad first attempts before understanding the mechanics of the test, users were allowed 2-3 minutes to get familiar with the setup and the Kinect. They could experiment with the 3D and understand the way the depth worked which many were unfamiliar with. They also got to listen to all the different sonification modes. This led to a more stable dataset which could be examined more easily.

All users tested all the different sonification modes, including the silent control mode. The users did 10 test runs (as described under Software) of each mode, for a total of 40 runs. To negate the effects of learning and order of appearance on the result, the order of the runs were shuffled individually for each participant. The reaction time of the participant were removed from the equation by showing a clear progress bar indicating when to start. The participant
Figure 3: All data gathered. Plotted with distance on the X-axis and time on the Y-axis. Since the average time increases as distance increases, we can see that Fitts law holds on our data set.

were also told ahead of starting each run which sonification mode were used in the current run. A complete user test took about 10 minutes from introduction to data on disk.

5 Results

This section handles the observations and analysis of the results. In total the experiments collected 480 points of data (12 participants with 40 runs each).

5.1 Data exclusion

To avoid invalid data caused by twitching from the Kinect some data points had to be removed from the analysed set. All runs with times exceeding this limit where discarded. This limit had to exclude possibly erroneous data but not exclude times that were just slow. To exclude as few real results as possible a limit of 10 seconds were chosen. This eliminated 16 data points and reduced our total amount to 464. It is noted that the eliminated data points were almost evenly distributed across the modes.

5.2 Data

In figure all the 464 data points are plotted with the sonification mode indicated by the colour and shape of the dot. From this plot it is hard to draw any conclusions since the different participants had so different times, and they also overlap each other. Therefore the analysis had to be done on a per user basis.
Figure 4: All user data sorted per user. Shows each users mean per sonification mode. The X-axis indicates which user and the Y-axis indicate the time. The figure clearly shows the spread of times between the users.

Due to the nature of the experiment (constant target size) and the movement style the participants used (people not moving their hand in a straight line and not completely knowing the direction and length of movement before moving), Fitts law does not apply directly to our data set. Some of its effects is however noticeable in this visualization of the data, mainly that increasing distance does increase the time taken. This can be seen in figure 3 where the average time goes up as distance goes up. There is some noise in the data, but this is expected since Fitts law does not give an absolute prediction, but should be treated more like an expected value.

When isolating all users and displaying their means separately as seen in figure 4 it is much easier to draw conclusions. Here each user is displayed so that their average time from each sonification mode is represented by a bar. The difference in average times between each user mentioned above can be easily spotted in figure 4.

To get a clearer view of how a participant reacted to sonification the means were normalized against the control mode, silent, as displayed in table 1. Here each row represents a user and their improvement relative to the silent control mode. The table is coloured to more easily get an overview of the information. Green cells represent a significant improvement of more than 5%, yellow cells has an improvement or decrease of less than 5% and orange cells represent user-mode combinations which have a performance decrease of more than 5%.

As can be shown in Table 1 the speed increase varied a lot between different participants. Some participants had speed increases of 20%, some had equally large slowdowns across all the audible modes compared to the silent mode.
<table>
<thead>
<tr>
<th>Frequency Δt%</th>
<th>Volume Δt%</th>
<th>Rate Δt%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-31%</td>
<td>-22%</td>
<td>-26%</td>
</tr>
<tr>
<td>29%</td>
<td>14%</td>
<td>14%</td>
</tr>
<tr>
<td>-6%</td>
<td>-20%</td>
<td>-2%</td>
</tr>
<tr>
<td>35%</td>
<td>35%</td>
<td>36%</td>
</tr>
<tr>
<td>-9%</td>
<td>-12%</td>
<td>6%</td>
</tr>
<tr>
<td>23%</td>
<td>3%</td>
<td>17%</td>
</tr>
<tr>
<td>-2%</td>
<td>-14%</td>
<td>-10%</td>
</tr>
<tr>
<td>-23%</td>
<td>-22%</td>
<td>-11%</td>
</tr>
<tr>
<td>4%</td>
<td>12%</td>
<td>-11%</td>
</tr>
<tr>
<td>-14%</td>
<td>-14%</td>
<td>5%</td>
</tr>
<tr>
<td>-15%</td>
<td>-13%</td>
<td>-9%</td>
</tr>
<tr>
<td>-16%</td>
<td>-21%</td>
<td>-37%</td>
</tr>
<tr>
<td>-3.9%</td>
<td>-7.1%</td>
<td>-3.5%</td>
</tr>
</tbody>
</table>
Figure 5: Confidence intervals (95%) for the complete data sets per mode (Shown in figure 3). The label indicates the mean value.

5.2.1 Normal distribution plot
Plotting the data as superimposed normal-distribution-plots for each mode as seen in figure 6 shows that the silent mode is below all the other modes throughout the spectrum. All the audible modes are grouped together, indicating that the nature of the sound is not very important, but rather that all types of sonification might be helpful during this type of motion. More directed research might be needed to find which mode is the best overall.

5.3 Observations
Pure observation of participants revealed that people with more experience of computers and especially computer games more quickly grasped the concept of navigating in 3D space. These participants also provided the shortest times.

Many participants reported that the biggest advantage with sonification was not the sound guiding you to the target, but rather the feedback of when you had arrived at the target. This feedback existed in all three audible modes. The large scale movements appeared to be more dictated by the visual feedback, while the small scale adjustments made when close to the target were more sound-assisted, especially when the first movements put the hand behind the target, in an angle that obscured the participants vision of the hand. When put in this situation with no sound feedback, some participants preferred to undo the motion until they regained visual feedback and tried again. These runs were among the slower recorded.

Also noted was that the rate mode has an easily distinguishable feedback
ceiling indicating that the target has been reached when the sound reaches a continuous tone. Both the frequency and volume are unbounded in their feedback ceilings, meaning that there are no discrete limits to indicate when the target has been reached. Instead the sound is muted when the target has been reached for these modes. This might have the effect that the user does not expect reaching the target in the same way he or she does when using rate sonification. Using a constant target volume or pitch would be much harder to discern from the build-up period since it does not differ that much from the previous feedback, in contrast to differentiating between a constant tone and an interrupted one, which is very easy.

5.4 Discussions

From our analysis there is no statistically significant correlation between sonification and improved times. There is however a trend between them which suggests that some improvement of performance is gained by adding sonification to user interaction.

The mode that showed the largest speed increase was the volume mode, but this was by almost all participants considered extremely annoying to listen to for extended periods, and might therefore not be the best idea for everyday or continuous-use applications. Some people are also quite sensitive to high volumes during computer use, and prolonged interaction might even cause hearing damage. This might instead favour the rate mode, which can be played at a much lower volume than the other modes and still be effective in conveying information to the user. This is due to the discrete nature of the sound, being either on or off, which takes a lot less thought to process than a continuous
change in frequency or volume.

It is very clear that participants reacted differently to sonification. Some of the participants even said that they got stressed by the sound, which worsened their performance. This type of reaction is probably linked to computer usage background. This was especially obvious when the participant had experience with 3D-games where positioning and sound awareness are important, such as FPS (First Person Shooter) games.

5.5 Error sources

During the test procedure there were a couple of possible sources of errors. One easily observable cause was the twitching from the Kinect sensor which was mentioned above. Noticeably, some participants experienced more twitching than others.

Another possible source for error was the inconsistency of the environment in which the tests were performed. Due to lack of funding we had to resort to using any available room either at the university or at one of the researcher’s home. This resulted in different brightness of the rooms and different backgrounds which might have affected the Kinect’s ability to reliably track the body.

Some participants were present during another persons test run which could affect the current participant both negatively and positively. It could also mean that their own performance were improved since they could observe the other person before their own attempt, and thus adjust their strategy.

The preparation of each participant could also have been more consistent. There were not a clear script of what to inform the participant about. This led to a few complaints when they further into the testing realised something that ought to have been explained beforehand. This in turn led to a varying amount of testing time before the actual test which could make some participants more used to the environment.

The data exclusion limit was perhaps chosen without statistical backing very early in the analysis process. Instead of a flat cut-off time, it might have been more prudent to only consider the middle 90% of data, or even the middle 90% of each participant’s data. This was however thought of very late in the analysis process. This would have resulted in 48 of the most extreme data points (both upper and lower extremities) being removed instead of, as it is now, 16. This might have provided better properties for statistical analysis on the data.

6 Conclusion

The gathered data supports the problem formulation, sonification does indeed improve performance of movement in 3D-space. The improvement is not massive, but there is a clear trend in the data. Almost all users experienced at least some speedup with audio feedback compared to the silent control mode.

Possibly the most important aspect of all the sonification modes was the confirmation when the user had reached the target and should hold still. This gave the user some much needed stimuli to react to since often the hand was obscured by the target when in close proximity. Some participants even suggested that this alone could help almost as much as other modes. If there is any truth to this would have to be a subject of another study.
Whilst the volume mode provided the biggest speed gain according to the user studies, it was also the most disturbing to listen to and required the most effort to use. A good compromise between effort and effect is rate sonification which also indicated a speed increase. Rate sonification also has the nice property of being volume-independent, i.e. it can be played at a constant and significantly lower volume than the other modes tested.

If the interaction is done over an extended period of time such as in a primary work task scenario with volume sonification there are also repeated strain injuries (RSI) to consider, where high volume under continuous periods can lead to hearing problems. The high volume requirement for the volume mode is especially disadvantageous when the sonification is not the only sound the user has to be able to hear. One example of such a situation might be a crane operator who needs to hear the foremans orders over the radio or a more ordinary office situation where the user only performs the interaction sporadically and wishes to have some music in the background or a video game with other sound effects in the background. All these cases are situations where rate sonification is probably the best suited mode. Frequency sonification, while not quite as ear-straining as volume mode, still suffers from the same volume-dependence when used. It also did not prove any better than the rate mode, which is why we do not recommend it.

Since people responded so differently to the feedback it is suggested that if implemented, it should be considered as an option in software, to be enabled as a personal preference on a user to user basis.
References


