KUNGL
TEKNISKA
HÖGSKOLAN

CID-112 • ISSN 1403-0721 • Department of Numerical Analysis and Computer Science • KTH

The quest for auditory direct manipulation: The sonified Towers of Hanoi

Fredrik Winberg & Sten-Olof Hellström

CID, CENTRE FOR USER ORIENTED IT DESIGN

Fredrik Winberg & Sten-Olof Hellström

The quest for auditory direct manipulation:The sonified Towers of Hanoi
Report number:CID-112
ISSN number: ISSN 1403-0721 (print) 1403-073X (Web/PDF)
Publication date: September 2000
E-mail of author: fredrikw@nada.kth.se, soh@nada.kth.se

Reports can be ordered from:

CID, Centre for User Oriented IT Design
NADA, Deptartment of Numerical Analysis and Computer Science
KTH (Royal Institute of Technology)
SE-100 44 Stockhom, Sweden
Telephone: + 46 (0) 8 790 91 00
Fax: + 46 (0) 8 790 90 99
E-mail: cid@nada.kth.se
URL: http://cid.nada.kth.se

The quest for auditory direct manipulation: The sonified Towers of Hanoi

Fredrik Winberg[1] and Sten Olof Hellström[2]

[1,2]Centre for User Oriented IT-Design, Royal Institute of Technology,
Lindstedtsvägen 5, SE-10044 Stockholm, SWEDEN

[2]Department of Music, City University,
London EC1V 0HB, UK

[1]fredrikw@nada.kth.se, [2]soh@nada.kth.se

ABSTRACT

This paper presents a study of an auditory version of the game Towers of Hanoi. The goal of this study was to investigate the nature of continuos presentation and what this could mean when implementing auditory direct manipulation. We also wanted to find out if it was possible to make an auditory interface that met the requirements of a direct manipulation interface. The results showed that it was indeed possible to implement auditory direct manipulation, but using Towers of Hanoi as the underlying model restricted the possibilities of scaling the auditory space. The results also showed that having a limited set of objects, the nature of continuos presentation was not as important as how to interact with the auditory space.

Keywords: Auditory interface, direct manipulation, sonification model, blind users

1. INTRODUCTION

The use of computers today is very dependent on the user's sight. The information is presented visually and sound is primarily used as very primitive queues for important visual information. This may not cause so much problems today for a blind computer user using a screen reader, given that all non-textual information has some sort of alternative description linked to it (which of course is not true, but for the sake of argument we will assume that this is so). But what about the other benefits that a graphical user interface gives a sighted user?

Representing the information using speech synthesis or Braille is a very linear way of presentation and has more in common with the old text based interfaces such as MS-DOS than it has with modern graphical user interfaces such as Windows or MacOS. And what about the next generation interfaces where the standard desktop environment is replaced by something completely different? Why should blind computer users still have to struggle with text based interaction?

1.1 Direct manipulation

Direct manipulation is a fundamental concept within HCI (human-computer interaction) and is based on the following properties:

- *Continuos representation of the object of interest.*
- *Physical actions or labelled button presses instead of complex syntax.*
- *Rapid incremental reversible operations whose impact on the object of interest is immediately visible.* (Schneiderman cited in Hutchins, Hollan, & Norman, 1985)

This means that you for example when moving a file instead of typing the command on your keyboard or choosing from a list of actions, you simply point your mouse at the file you want to move, grab it by pressing down the mouse button, drag it to the place you want it to be and drop it by releasing the button. Another important feature of direct manipulation is that it relies on recognition rather than recall, for example the use of menus helps the user to remember the name instead of forcing the user to memorise the exact name and the exact syntax of the command of interest.

Direct manipulation has been very influential in today's graphical user interfaces and will influence the way we interact with computers for a long time.

Paper presented at the 3rd International Conference on Disability, Virtual Reality and Associated Technologies ICDVRAT 2000, Sardinia, Italy, 23-25 September 2000

1.2 Screen readers

In present screen readers for blind computer users, direct manipulation as well as a number of other important features of the graphical user interface are missing. Mynatt has summarised five goals for screen reader interface design (1997):

- Access to functionality. The screen reader should at least give the user access to the same functions as are presented in the graphical user interface. In a graphical user interface, most functions are represented as pull-down menus. The screen reader should give the blind computer user access to this functionality.
- Iconic representation of interface objects. The screen reader has to be able to recognise and present the same information as is communicated by the visual appearance of the interface objects such as the picture, size and colour. For example, the picture of a trash can on an icon in MacOS symbolises that the icon represents the trash can and that it is a suitable place to throw things one want to get rid of. The shape of the icon tells the user whether the trash can is empty or has things in it.
- Spatial arrangement. The spatial arrangement of the graphical objects also conveys information that helps the user in structuring and working with many tasks at once. The screen reader should offer this functionality.
- Constant or persistent presentation. Visual information is not time dependent in the same way as audio is. The visual information exists in physical space and can be obtained and reviewed at any time; this is not the case for audio information. The screen reader should support this kind of temporal independence.
- Direct manipulation. The screen reader should give the user the same powerful means of interaction as direct manipulation does.

If auditory direct manipulation is to be implemented, all of these items have to be solved.

Auditory direct manipulation is a rather uncharted territory both in research and development, given that we talk about real direct manipulation and not just interacting directly or almost directly with interface objects. In the GUIB project for example (GUIB Consortium, 1995) the work has been concerned with giving the blind computer user a more direct way of interacting with interface objects, but it has not dealt with direct manipulation itself. Other work has been done on complex auditory interfaces (see for example Gaver, Smith, & O'Shea, 1991), but most of these has been monitoring tasks were the focus has been on the display of information rather than the interaction with it (Saue, 2000).

2. GENERAL GOALS

The two questions we want to address are

- Is auditory direct manipulation at all possible?
- Is auditory direct manipulation at all interesting or do we have to seek other paradigms for interaction with an auditory interface?

In order to answer the above questions we have implemented three different audio-only, non-visual, versions of the game "Towers of Hanoi" (see for example Ball, 1939). We also performed two user studies on these three versions. The goal of the studies was to investigate the first principle of direct manipulation, continuos presentation, and what this could mean in an auditory interface.

The three different levels of continuos presentation under study are *parallel*, *serial*, and *overlapping* presentation mode. The first extreme case is when all sounds keep repeating simultaneously, the parallel presentation. The other extreme is when there is no overlap at all and the sounds are played in sequence, the serial presentation. Finally, we implemented a mixture of theses two with a slight overlap, the overlapping presentation (see Sonification model for a more detailed description of these).

3. TOWERS OF HANOI

The game "Towers of Hanoi" consists of three towers where a number of differently sized discs are placed. Initially all discs are placed on the leftmost tower with the discs placed in order with respect to size with the smallest disc on top. The goal is to move all the discs to the rightmost tower. You are only allowed to move one disc at a time and this disc has to be on top of a tower. You can move this disc to any of the three towers just as long as you don't move a larger disc in top of a smaller one. This game can be played with as many discs as you want without having to use more than three towers. The number of moves to complete the game increases rapidly when adding discs, the number of moves it takes to solve the game for n discs is 2^n-1. This means that three discs take 7 moves, eight discs takes 255 moves and sixty-four discs takes $1.8 \cdot 10^{19}$ moves to complete.

We chose this game for three reasons; (1) we wanted to have a game that could be fun and challenging to solve, not a typical experimental task, (2) the rules of the game are fairly easy to learn and the strategy is straightforward and doesn't change when increasing the number of discs, just the number of steps in the solution path, and (3) it's easy to show the subjects a wooden model of the game in order for them to learn how to solve the game (this applies both for blind and sighted subjects).

4. SONIFICATION MODEL

In the experiment we studied two factors, *game complexity* and *presentation mode*. Game complexity varied at two levels, referred to as *3disc* and *4disc*. Presentation mode varied at three levels referred to as *serial*, *overlapping* and *parallel*. See the next section for a discussion of the experimental design.

In the 3disc condition, three discs were moved between three towers. In the 4disc condition, four discs were moved between three towers. We also want to represent the height of any given disc on a tower. This requires the representation of up to four discs, three horizontal locations and up to four vertical locations. To accomplish this in sound, each of the discs is identified through associating it with a sound of a specific quality, and the positions of the discs are given through spatialising the sounds in stereo, varying their amplitude envelopes and varying their length. We discuss these features of *disc identity* and *disc location* below.

4.1 Disc identity

Timbre and pitch variations are used to individuate the discs. The larger the disc, the lower the pitch. Let us call the largest disc 1, and the smallest disc 4. The sounds are mistuned with respect to each other and only rarely have partials of the same frequencies, which helps to maximise their discriminability. The fundamental frequencies of the sounds are: 118 Hz (disc 1), 181 Hz (disc 2), 336 Hz (disc 3) and 456 Hz (disc 4). (In the 3disc condition, only discs 1, 2 and 3 were used.)

Disc 1 has a sparser harmonic spectrum than disc 2 and, similarly, disc 3 has a sparser spectrum than disc 4. Furthermore, discs 1 and 2 have fewer high frequency partials than discs 3 and 4. Any combination of discs will differ from any other combination in terms of both pitch and timbre, and do so in a unique way. There is a large gap in frequency between disc 2 and 3 so as to stop the sounds from fusing together when three or more are heard from the same location in the stereo image (cf. Bregman, 1990).

The distinctions between the discs involve some redundancy or overcoding, being conveyed through simultaneous variations in more than one auditory dimension. This is necessary when only one single auditory dimension is difficult to perceive in a complex auditory space (Kramer, 1994).

4.2 Disc location

To represent the tower a disc is located on, stereo panning is varied, left, centre and right stereo locations are used. The spatial discriminability of the sounds is further enhanced by varying their amplitude envelope. Individual discs are presented by pulsing their sounds. The character of the envelope of each pulse is varied to indicate which tower a disc is located on. If a disc is placed on the left or right tower, the percentage ratio between attack and decay is 0:100. If a disc is placed on the middle tower, the same ratio between attack and decay is 50:50.

As with disc identity, the spatial location of the discs is represented redundantly by simultaneously varying panning and amplitude envelopes.

A disc's vertical position is represented by the length of the pulse, the higher the disc is placed the shorter the sound. For example, if two discs are placed on the same tower, the one in the lowest position has a longer pulse length than the one on top. The pulse lengths are 900, 600, 333, and 238 ms.

4.3 Presentation modes

To represent the overall configuration of the Towers of Hanoi at any given moment, three (in 3disc) and four (in 4disc) inter-related series of pulses are to be heard. The relative timings of these series, and the inter-pulse intervals within them, have been designed in three different ways.

- The serial condition. The pulses for the discs are repeated in numerical order without any delay or overlap. As the pulses vary in length to represent the height on the tower, the inter-pulse interval in this condition will vary depending on the location of the discs.

- The overlapping condition. The inter-pulse onset interval is set to a constant value of 300 ms. Accordingly, a pulse associated with a disc will overlap with that of another if the following disc is not placed on top of three others (it's pulse length is smaller than the inter-pulse interval). Discs 1, 2 and 3 (and then 4 in 4disc) are repeatedly pulsed in order.
- The parallel condition. The discs are pulsed continually and simultaneously. For each disc, the onset of a new pulse occurs immediately after the release of the previous one.

4.4 Mouse location

We used the mouse as the input device. In order for the user to track the mouse cursor, the amplitude of the discs on the tower where the cursor is located on is increased while the other discs amplitudes are decreased (the difference is 1:3). Just using this amplitude focus can cause problems when all discs are located to either right or left, since there are no sounds to indicate the difference between middle and the opposite side. To solve this problem we are also using transition tones that will sound when moving the cursor from one tower to another. If moving to left or right from the middle, a short high tone (600 Hz for 500 ms) will sound from left or right. If moving from left or right to the middle, a short lower tone (400 Hz for 500 ms) will sound from both left and right.

5. THE STUDY

5.1 Hypotheses

- It is possible to design an auditory interface that meets the requirements of direct manipulation as defined above.
- The overlapping presentation mode will be the version that most subjects will both prefer and get the best results from using. This will be further emphasised when increasing the complexity (using four instead of three discs).

When presenting the sounds using the parallel presentation mode, it will be easy to get a general overview of all objects, but the separation could be quite hard when having many objects. Furthermore, continuos sounds, or continuos presentation of sounds, could be harder to separate (cf. Gaver, Smith, & O'Shea, 1991) and masking would be more likely.

When using the serial presentation mode there is no problem of separation of the objects since there is no sounds ever overlapping. On the other hand, the general overview is harder since the user has to wait until all sounds has been played to get an overview. If the auditory interface is supposed to support direct manipulation and the number of objects is large, this is could hardly be called continuos presentation in that case.

Since both of these presentation modes both have drawbacks and advantages in comparison to one another, a combination of these seems to be the most appropriate, namely the overlapping presentation mode.

5.2 Experimental design

The experiment is designed to be a two factor within subject design. The first factor is presentation mode and is varied at three levels, serial, overlapping and parallel. The second factor is game complexity and is varied at two levels, three and four discs. Each subject played the game once for every combination of the two factors, which means that every subject, played the game six times. The sequence of the combinations was counterbalanced using a Latin square.

During the experiment two quantitative measurements were made, number of errors (or rather the number of extra steps in the solution path compared with the optimal path), and time to complete. After the session the subject answered questions about which presentation mode they preferred and which they thought they performed best with.

The quantitative data had to be analysed using nonparametric statistical methods, since the measurements neither could be classified as ratio or interval, but rather as ordinal measurements. Additionally, these methods are very insensitive to extreme values, something that is important in an experiment were one might expect a learning effect that will vary between different subjects. The three level factor (presentation mode) was analysed using the Friedman two-way analysis of variance by ranks. The two level factor (game complexity) was analysed using the Wilcoxon signed ranks test.

The experimental set-up was very simple. The subject used a pair of earphones and a regular computer mouse, the computer screen was turned away from the subject and was used exclusively by the session leader to monitor what the subject was doing.

A session started with the subject being informed about what was going to happen during the experiment and the purpose of the study. After this, the subject learnt to play the game using a wooden model of the game. This continued until the subject knew how to solve for both three and four discs without making any errors. By doing this we are trying

to even out differences in prior knowledge of the game and get all subjects to have a useful and similar model in mind when solving the auditory version of the game. After the subject has been accustomed to the game, the wooden game is taken away. Now the sonification model is presented. All aspects are described and demonstrated to the subject and the subject is allowed to ask questions and hear every detail as many times as he or she wants. The subject is also informed that this is the last chance of asking any questions about the game or the sonification model. When the subject thinks that he or she knows the sonification model the experiment starts. After all combinations of the game has been completed, the session ends with the subject answering a number of questions about preferences and perceived performance

5.3 The first study

The results of the first study was more concerned with the sonification model and the experimental design than on the question of continuos presentation (Winberg & Hellström, 2000). Of the twelve subjects, three could not complete all or some of the conditions. The two things that caused these dropouts, and caused problems for all subjects for that matter, were the mouse interaction and the instructions.

The sonification model during this first study differed from the one described above in how the mouse interaction was designed. The amplitude ratio was smaller (1:2) and there were no transition tones. Most subjects had big problems tracking the mouse cursor when there were no transition tones. Many subjects found it very hard to find the middle tower, flipping the cursor from left to right without ever finding the middle. This had a very randomised effect on their results, making the collected data very unreliable.

The instructions that the subjects received were also something that caused problems. Many subjects simply did not understand the sonification model at all, invalidating the basic assumption that all subjects would have a model of the game and an understanding of the sonification model before the experiment started.

Despite all these problems encountered during this first study, as stated above nine out of the twelve subjects actually understood and achieved well in the experiment. Due to these problems, we had no interesting or valid data to do any statistical calculations on. But the qualitative results pointed towards a strong preference for the overlapping version and the subjects also thought that they performed better using this presentation mode, even though this was nothing that could be deduced from the measured data.

5.4 The second study

When planning the second study, we introduced transition tones and increased the amplitude difference to enhance the mouse interaction (see Mouse location above). We also changed the instructions and described the sonification model more thoroughly to the subjects. By doing these two adjustments, we hoped that that the problems encountered in the first study would be eliminated. The rest of the set-up and experimental design remained the same in the second study.

The outcome of the second study was very encouraging. None of the problems with the mouse interaction from the first study appeared, none of the subjects had any problems tracking the mouse cursor or finding the middle tower.

Again, as suspected, the subjects preferred the overlapping version, but when it came to the statistical analysis, the results were very surprising. The hypothesis that the overlapping version would be better could not be supported by the analysis. The only significant results we got were that when using the overlapping presentation mode it took more time to complete with four than three discs (Wilcoxon z_{time}=-2.275, $p<0.05$), a result that is not interesting at all since the number of moves is greater. The differences in number of errors between presentation modes were also significant, but only when using three discs (Friedman χ^2_{errors}=7.032, $p<0.05$), not when using four discs (Friedman χ^2_{errors} =0.391, $p>0.05$, χ^2_{time} =1,167, $p>0.05$), something that the hypothesis suggested. We did not find any significant difference between the three presentation modes in general either (not taking the number of discs into account) (Friedman χ^2_{errors} =2.909, $p>0.05$, χ^2_{time} =3,583, $p>0.05$).

5.5 Subjects

We used twelve subjects in each study. Of these, just one was blind in the first study and none during the second. The reason for not using more blind people, who indeed are the focus of this work, is that in this early stage of this work we wanted to fine tune the experiment and the sonification model so that we wouldn't "waste" the blind subjects by letting them participate in an experiment that might not give any interesting or valid results. Additionally, in this early stage we are interested in such a low level questions that the difference in experience of auditory interfaces between blind and sighted subjects is of minor importance. The continuation of this work will definitely involve blind users in defining, designing and evaluating the auditory direct manipulation interfaces.

6. DISCUSSION

The first question one might ask is whether a sonification of the game Towers of Hanoi is complex in a relevant and interesting way and if it really helps answering the hypotheses. The sonification and the study of Towers of Hanoi as an auditory direct manipulation interface should not be seen in isolation. As will be pointed out below, this is just the first of a number of studies on auditory direct manipulation. However, this study has given us important pointers were to go from here.

The second question is why we have chosen the sounds that we have, why haven't we chosen real world sounds like auditory icons (Gaver, 1994) and used natural mappings. The reason for choosing abstract sounds is primarily the scalability (cf. earcons in Blattner, Sumikawa, Greenberg, 1989) of the model. Adding to this is the fact that the concept of metaphorical or real world sounds is hard when displaying abstract concepts. There is even research that reports that there are many acoustical mappings of auditory variables that doesn't necessarily have to be perceived the same by all listeners, even though they would be considered intuitive or natural (Walker, Kramer, & Lane, 2000).

The sonification of "Towers of Hanoi" presented in this paper meet the requirements of a direct manipulation interface.

- The objects are presented in a continuos manner (or rather in three different ways that all could be considered continuos enough). This means that the user rather than scanning the whole interface for objects with the mouse risking missing some of them can hear all objects all the time without actually "looking" for them.
- We have physical actions instead of complex syntax. Instead of choosing the appropriate action from a menu or typing it on the keyboard, the user picks up, moves and drops the object just like a graphical user interface.
- All actions are rapid, incremental and reversible with immediate feedback. There are no detours when performing an action, the only way of moving a disc is using the shortest path between the start and the goal. If the user decides that a move is wrong, it is as easy to move it back as it was moving it there.

The hypothesis that the overlapping presentation mode would be better could not be supported by this study, even though this is what we expected. We have three different explanations to this.

- Since the game Towers of Hanoi in itself can be complex and hard to solve for some people, the complexity or number of auditory objects that is possible to present is limited by the underlying model. During an early pilot study, we concluded that subjects could have problems solving the game if we used five or more discs. Therefore, we had to limit the number of objects to four. When increasing the complexity of the auditory space we might get different results. This calls for more studies of these kind of auditory interfaces where the complexity is not limited by the underlying model but rather by the limits of the sonification model and the users perception.
- Using the mouse interaction with amplitude focus facilitates the interaction with a complex auditory space. When having a limited set of auditory objects the means of interaction is more important than the way to solve continuos presentation. The way we implemented the mouse interaction increases the amount of objects and the complexity that is possible to interact with.
- The sonification model used is robust enough from being influenced by presentation mode, at least in this specific context. Again, this calls for further studies using other contexts in order to assess the validity of this specific approach

All these three explanations call for further studies. The final stage of this study of Towers of Hanoi is to make an extensive case study with blind subjects were the qualitative aspects of this auditory interface is investigated. This third study will help us understanding more about auditory direct manipulation, how it could be used and what type of applications would be interesting to implement using this paradigm.

We believe that auditory direct manipulation is indeed both possible and a promising way of interacting with an auditory interface. It will provide blind computer users with a completely new way of interacting with a computer, a way that so far has been inaccessible.

7. REFERENCES

Ball, W. W. R. (1939). *Mathematical recreations & essays* (11th ed.) (pp. 303-305). London: Macmillan & Co.

Blattner, M., Sumikawa, D., & Greenberg, R. (1989). Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction, 4*(1), 11-44.

Bregman, A. S. (1990). *Auditory scene analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press.

Gaver, W. W. (1994). Using and Creating Auditory Icons. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 417-446). Reading, USA: Addison-Wesley.

Gaver, W.W., Smith, R.B., & O'Shea, T. (1991). Effective sounds in complex systems: The ARKola simulation. In *Proceedings of CHI'91* (pp. 85-90). New York: ACM.

GUIB Consortium. (1995). *Final Report of the GUIB Project: Textual and Graphical Interfaces for Blind People*. London: Royal National Institute for the Blind.

Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1985). Direct manipulation interfaces. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 87-124). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kramer, G. (1994). Some Organizing Principles for Representing Data with Sound. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 185-221). Reading, USA: Addison-Wesley.

Mynatt, E. D. (1997). Transforming graphical interfaces into auditory interfaces for blind users. *Human-Computer Interaction, 12*, 7-45.

Saue, S. (2000). A model for interaction in exploratory sonification displays. In *Proceedings of ICAD 2000* [online proceedings]. URL http://www.icad.org/websiteV2.0/Conferences/ICAD2000/ICAD2000.html (visited 2000, July 11).

Walker, B.N., Kramer, G., & Lane, D.M. (2000). Psychophysical Scaling of Sonification Mappings. In *Proceedings of ICAD 2000* [online proceedings]. URL http://www.icad.org/websiteV2.0/Conferences/ICAD2000/ICAD2000.html (visited 2000, July 11).

Winberg, F. & Hellström, S. O. (2000). Investigating Auditory Direct Manipulation: Sonifying the Towers of Hanoi. In *CHI 2000 Extended Abstracts* (pp. 281-282). New York: ACM.