UNIVERSITY
of
GLASGOW

Deictic Spatial Audio Target Acquisition in the Frontal Horizontal Plane

Georgios Marentakis

Submitted for the degree of Doctor of Philosophy

Abstract

The application of direct manipulation principles in the audio domain, by taking advantage of the directional nature of hearing and contemporary virtual audio systems is a design proposal that could be applied in a variety of application areas including virtual and augmented reality, mobile Human Computer Interaction and also applications for visually impaired people.

However, no research has been done on assessing the feasibility of the audio modality for supporting the deictic or pointing based interactions common in direct manipulation interfaces. The success of a direct manipulation environment is dependent on fast and accurate pointing, however it is questionable whether our hearing sense can support the acquisition of targets in the way vision does. The research questions the thesis investigates are thus defined to be:

RQ 1	How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?
RQ 2	What are the factors that affect deictic spatial audio target acquisition?
RQ 3	How can we evaluate the usability of deictic spatial audio target acquisition?

The Research Questions are answered using existing literature and by performing experimental investigations. It is found in the literature that the lower spatial resolution of hearing does not provide detailed support for primary and secondary submovements in the way vision does. The problem is amplified by limitations in the fidelity of commercial spatial audio reproduction systems.

The thesis proceeds by examining the feasibility of spatial audio target acquisition in an experimental context. It is found that feedback marked targets areas are necessary for participants to successfully point to spatial audio targets. In addition, it is found that the gesture used affects the selection effectiveness of spatial audio target acquisition. When selecting a feedback marked spatial audio element it is found that targets should be associated with a width of 9°, 16° and 18.5° depending respectively on whether users use a touch tablet, their hands or their heads to select to enable 70.7% selection success rate.

The second experiment shows that the type of feedback used to enhance localization ability affects the efficiency and the effectiveness of the interaction. Feedback marked audio elements are found to be the best solution compared to loudness or real time orientation update cues.

The third experiment proposes and evaluates a spatial audio target acquisition task which is found to be usable and insensitive to bone conductance presentation. Monaural presentation is found to degrade interaction speed by a factor of 2 due to the lost localization cues. In addition, it is found that under headphone presentation the task of acquiring a feedback marked spatial audio element is highly correlated

with Fitts' law and thus the techniques used for the evaluation of visual target acquisition can be applied in the spatial audio domain.

In the fourth experiment, this result is verified and a detailed quantification of the effect of target width and the ratio of distance to target to target width is obtained. In addition, the selection task is found to be usable when mobile, however a degradation of performance in the order of 20% is observed. Feedback marking the display elements is found not to affect perceive workload or percentage preferred walking speed. Finally, when feedback is removed that acquisition task is found to be prone to serious performance problems mostly related to the accuracy of selections.

Based on the experiments and the theoretical analysis the thesis contributes by disambiguating the task of selecting a spatial audio target and providing estimators of display parameters and a number of guidelines for designers. In addition, a theoretical framework inspired from visual target acquisition is found to be appropriate for the evaluation of the acquisition of feedback marked spatial audio targets thus setting the foundation for future investigations in gesture interaction with spatial audio displays.

Acknowledgements

I would like to acknowledge the support of my supervisor Stephen Brewster in all aspects of the development of the thesis. I would also like to thank him for allowing me to freely explore the research field my thesis was associated with. I would also like to thank my second supervisor Roderick Murray-Smith for his support and useful discussions. Finally, I would like to thank the members of the Multimodal Interaction Laboratory for their collaboration and help in numerous aspects of the thesis development. Thanks to Lorna Brown, Steven Wall and Andy Crossan for reviewing parts of this document.

Special thanks should also be given to the people that supported me in an emotional way throughout these three years, who are too many to mention in this paragraph. Thanks to all of my friends and loved ones for being there for me. The list obviously starts with my parents and my brother. Special thanks to Jacqueline Hall, Nick Chozos, Anna Hatzidaki and Vangelis Mitrou.

Finally, I would like to thank Adrian David Cheok for giving me the opportunity to work for three months in the exciting environment of the Interaction and Entertainment Research Centre in Singapore.

And a small thing for me to remember, the PhD was a goal and a trip and as in all worth while trips it was necessary to wander and to trust. I am particularly happy that the process of completing my PhD taught me many small and big things and for the fact that my PhD was a fruitful, rewarding and demanding companion.

Declaration

The contents of Chapters 5, 6, 7 and 8 have been published as

- Marentakis, G. and Brewster, S.A. *A Study on Gestural Interaction with a 3D Audio Display*. In Proceedings of MobileHCI2004 (Glasgow, Scotland), Springer LNCS Vol. 3160, pp. 180-191.
- Marentakis, G. and Brewster, S.A. *A Comparison of Feedback Cues for Enhancing Pointing Efficiency in Interaction with Spatial Audio Displays*. In Proceedings of MobileHCI 2005 (Saltzburg, Austria). ACM Press, pp. 55-62
- Marentakis, G. and Brewster, S.A. *Effects of Reproduction Equipment on Interaction with a Spatial Audio Interface.* In Vol II Proceedings of ACM CHI 2005 (Portland, Oregon, USA). ACM Press, pp. 1625-1628.
- Marentakis, G. and Brewster, S.A. *Gesture Interaction with Spatial Audio Displays: Effects of Target Size and Inter-Target Separation.* In Proceedings of ICAD2005 (Limerick, Ireland), July 2005. ICAD, pp. 77-84.
- Marentakis, G. and Brewster, S.A. *Effects of Feedback, Mobility and Index of Difficulty on Deictic Spatial Audio Target Acquisition in the Horizontal Plane.* To appear in Proceedings of ACM CHI 2006 (Montreal, Canada), ACM Press Addison-Wesley

The contents of thesis are however associated only with the author's personal work.

Table of Contents

1 Introduction 1

1.1 Motivation & Aims 4

1.2 Research Questions 5

1.3 Thesis Walkthrough 5

2 Spatial Sound Perception in Real & Virtual Environments 7

2.1 Introduction 7

2.2 Theory of Sound Localization 7

2.3 Spatial Sound Perception in Real Environments 11

2.3.1 Notes on the Perception of Direction 11

2.3.2 Notes on the perception of distance 15

2.3.3 Cross Modal Effects 17

2.3.4 Perception of Auditory Motion 18

2.4 Spatial Audio Reproduction in Virtual Environments 19

2.4.1 Details on HRTF Estimation and Implementation 19

2.4.2 Sound Localization in systems using HRTF filtering 21

2.5 Discussion 27

2.6 Conclusions 28

3 A Review of Existing Research in Spatial Audio Displays 30

3.1 Introduction 30

3.2 Review of Application Designs based on spatial audio 31

3.2.1 Handy Sound & MAW 31

3.2.2 Spatial Audio & Teleconferencing 32

3.2.3 Head Mounted Spatial Audio Pie Menus 33

3.2.4 Grid Menu 33

3.2.5 AudioStreamer 34

3.2.6 Dynamic Soundscape 34

3.2.7 Audio Hallway 35

3.2.8 Nomadic Radio 35
3.2.9 Virtual Audio Guidance and Alert System for Commercial Aircraft Operations 36
3.2.10 3D Web Browser & WIRE 36
3.2.11 Active Localization of Virtual Sounds 38
3.2.12 Audio GPS 38
3.2.13 GPS Tunes 38
3.2.14 Diary in the Sky 39
3.2.15 Monitoring Background Activities (Spatial Audio Progress Bar) 39
3.2.16 AudioDoom 39

3.3 Summary of the Design Choices in Spatial Audio Displays 40

3.3.1 The Effect of choices on the Spatial Parameters of the Display Elements 41
3.3.2 Egocentric vs. Exocentric Designs 42

3.4 Reproduction 44

3.5 Control 44

3.5.1 Gesture Control 44
3.5.2 Control through a Speech Recognized Command Vocabulary 45
3.5.3 The Control Option chosen in the Thesis 45
3.5.4 Tracking Technology Alternatives 46

3.6 Application Areas 48

3.7 Implementation Issues 50

3.8 Identifying the requirements for supporting direct manipulation 51

3.8.1 Auditory Direct Manipulation 53
3.8.2 Auditory Signs 54

3.9 Discussion 56

3.10 Conclusions 58

4 Evaluating & Modelling Pointing Interactions 59

4.1 Introduction 59

4.2 Empirically Derived Models 59

4.2.1 Effective Target Width 61
4.2.2 Logarithmic Models 63
4.2.3 Linear Models 65

4.2.4 Power Models 67

4.3 Theoretical Investigations & the Speed-Accuracy Trade-Off 67

4.3.1 The information theoretic perspective 68

4.3.2 The iterative corrections model 69

4.3.3 The stochastic optimized submovements model 70

4.4 Applications in Human Computer Interaction 73

4.4.1 Applying Fitts Law to two and three dimensional targets 75

4.5 Comments on the models, their scope and their theoretical backgrounds . 76

4.6 Conclusions 79

5 An initial investigation into spatial audio target acquisition 81

5.1 Introduction 81

5.2 Rationale 81

5.3 Adaptive Psychophysical Methods 83

5.4 Experiment 85

5.5 Stimuli and Apparatus 86

5.6 Experimental Design & Hypotheses 87

5.6.1 Hypotheses 87

5.7 Experimental Task 87

5.8 Procedure & Participants 88

5.9 Results 88

5.10 Discussion 93

5.11 Conclusions, Guidelines & Future Directions 94

6 An investigation into deictic interaction in egocentric and exocentric displays. The effect of feedback cues and distracter sounds. 96

6.1 Introduction 96

6.2 Rationale 96

6.3 An introduction to the issues associated with simultaneous presentation of audio streams 97

6.4 Feedback Cues 98

6.5 Evaluation Methodology 100

6.6 Experiment 101

6.7 Experiment Design & Hypotheses 101

6.7.1 Experimental Hypotheses 102

6.8 Stimuli & Apparatus 102

6.9 Experimental Task 104

6.10 Procedure & Participants 105

6.11 Results 105

6.11.1 Time Analysis 105

6.11.2 Accuracy Analysis, Throughput and Effective Target Widths 107

6.11.3 Workload Analysis 110

6.12 Discussion 110

6.13 Conclusions, Guidelines & Future Directions 113

7 An investigation on the effects of Reproduction Equipment, Target Size and Inter-Target Separation on Gesture Interaction with a Spatial Audio Display 115

7.1 Introduction 115

7.2 Rationale 115

7.3 Background on the examined reproduction techniques 116

7.4 Background on the examined display segmentation techniques 118

7.5 Experimental Design 120

7.5.1 Experimental Hypotheses 120

7.6 Experimental Task 121

7.7 Stimuli & Apparatus 121

7.8 Procedure & Participants 122

7.9 Results 122
7.9.1 Time Analysis 122
7.9.2 Accuracy & Workload Analysis 123
7.9.3 Additional Observations 125
7.10 Discussion 130
7.10.1 Discussion on the Effects of Reproduction Equipment 130
7.10.2 Discussion on the Effects of Display Segmentation and Interaction Patterns 131
7.11 Conclusions, Guidelines and Future Directions 134
8 An Investigation into the effects of Mobility, Feedback and Index of Difficulty on Spatial Audio Target Acquisition in the Frontal Horizontal Plane 137
8.1 Introduction 137
8.2 Rationale 137
8.3 Mobility 138
8.4 Experiment Design & Hypotheses 139
8.4.1 Target Width and Distance to Target Manipulation 140
8.5 Experimental Task 141
8.6 Stimuli & Apparatus 142
8.7 Procedure & Participants 143
8.8 Results 143
8.8.1 Performance with on-target feedback 144
8.8.2 Performance without feedback 147
8.9 Walking Speed Analysis 148
8.10 Workload Analysis 148
8.11 Can Fitts' law be used to describe 3D audio target acquisition? 149
8.12 Discussion 151
8.13 Conclusions and Guidelines 152
9 Conclusions 154

9.1 Introduction **154**

9.2 Summary of the Thesis **154**

9.3 Thesis Contributions **157**

9.4 Thesis Limitations & Future Directions **160**

9.5 Future Research **162**

10 APPENDIX ***165***

11 References ***169***

List of Figures

Figure 1.An illustration of azimuth, elevation and distance, the variables that are used to define the position of a sound source relative to a listener's head. 0° of azimuth corresponds roughly to the direction of a user's nose. The point of reference is assumed to be in the centre of the head, adapted from [61]. 8

Figure 2. Illustration of the median, frontal and horizontal frames. These areas are commonly used to refer to the location of sounds, relative to the head of a person due to the fact that they result in certain physical properties for the signal that reaches the ear, adapted from [61]. 9

Figure 3. An illustration of the cone of confusion. Sounds located around the cone result in the same interaural time and intensity differences, and it is thought that this the reason that their position cannot be reliably identified in the absence of pinnae cues, adapted from [97]. 10

Figure 4. Localization error at a number of angles, adapted from [14]. Mean perceived direction as well as its deviation are illustrated for sounds originating at the direction of the arrow. 12

Figure 5. The minimum audible angle (MAA) for sinusoidal signals, plotted as a function of frequency, each curve shows results for a different reference direction. As can be seen for certain directions of a sound, the minimum audible angles can vary a lot as a function of frequency, adapted from [85]. 13

Figure 6 Illustration of the concept of directional bands, adapted from [14]. H stands for behind, v for forwards. Sounds containing frequency components inside the bands are perceived as coming from backwards or forward, irrespective of their actual direction. 14

Figure 7, Perception of distance vs. the actual distance from the source, adapted from [14]. If an one to one relationship between actual and perceived distance the result would be a straight line with a slope of 45° 16

Figure 8. The acoustically open transducer used by Langendijk and Bronkhorst. The trnsducer is positioned at a distance to the ear to allow for a microphone to be placed in the ear. 22

Figure 9. Localization accuracy as a function of angle in front of the user, 0 and 180 correspond to extreme left and extreme right. It is seen that people tend to perceive sounds as coming from the sides even when they originate from diagonal directions, adapted from [48]. 37

Figure 10. The Polhemus FastTrack System 46

Figure 11. The Intersense Intertrax tracker 47

Figure 12. The P5 glove 47

Figure 13. The MT-9B tracker. 48

Figure 14. Fitts' Experimental Design, a participant selecting two targets in a discrete fashion. Participant is shown resting in the home position waiting for a signal to start moving to the indicated direction (adapted from [37]). Participants return to the home position after selection. When participants alternate between the targets continuously, the task is called continuous. 60

Figure 15. Illustration of the trajectory encountered when a user moves the a virtual pointer through a constrained path, of length A and width W. A linear model has been found to predict movement time in such a case. 66
Figure 16. Outline of the assumptions of the deterministic iterative-correction model with respect to movement trajectory. The horizontal axis represents distance and the vertical axis velocity. The curves correspond to successive submovements between an initial home position and a target region. 71
Figure 17. Selection from a bulls eye menu as the one used by Friedlander *et al*. Adapted from [39]. 75
Figure 18. A graph showing movement time predictions for the competing models, for model values of intercept = 0.5 and slope = 0.28. 77
Figure 19. A graph showing Index of Difficulty predictions for the competing models, for model values of intercept = 0.5 and slope = 0.28. 78
Figure 20. Selecting a virtual 3D audio source in the hand pointing condition. 87
Figure 21. Effective selection angle as a function of sound direction for each interaction technique. 89
Figure 22.Mean absolute deviation from target and its standard deviation versus sound direction and interaction technique. 91
Figure 23. Mean ease of use ratings for each interaction technique. 92
Figure 24. Mean Comfort ratings for each interaction technique 92
Figure 25. Experimental Task, Visualization of the hand of a participant alternating between the two targets while tested in the display. 103
Figure 26. Performing the task in Experiment 2. A user is illustrated selecting a 3D audio source. 104
Figure 27. Mean Times to Complete Trials as a function of the number of the display elements for each feedback type used in the experiment. 106
Figure 28. Mean absolute deviation from target and its standard deviation means calculated using data from all cases of the number of display elements factor. 107
Figure 29. Effective target widths for the different feedback cues calculated across all display populations. 109
Figure 30. Mean Throughput for the different feedback cues calculated across all display populations. 112
Figure 31.The spatial audio acquisition task under examination. A is the distance to target and W the target width. 116
Figure 32. The different headphone types used in the experiment, bottom left is the single earpiece (Panasonic RP-HS50), top left are the bone conductance headphones (Vonia EZ – 3200P) and to the right are the Sennheiser HD 200 headphones. 118
Figure 33 Mean time and standard deviation for the three presentation methods the two display designs and associated distance paths. The first line in the legend corresponds to the distances in the minimal display, while the second in the distances in the maximal display. 123

Figure 34. Mean time spent in overshooting the target per display type, headphone case. 125
Figure 35. Mean time spent in overshooting the target in the case of bone conductance headphones 126
Figure 36. Histograms of selection angles for all target positions, headphone presentation. 126
Figure 37. Histograms of selection angles for all target positions, bone conductance presentation 127
Figure 38. Histograms of selection angles for all target positions, monaural presentation 128
Figure 39. A participant selecting a virtual 3D sound source while walking and wearing the experimental apparatus. 142
Figure 40. Means and standard deviations of the time and accuracy scores for standing and mobile participants who did and did not receive feedback 144
Figure 41. Mean time scores for participants who received on-target feedback as a function of ID and target width. 145
Figure 42. Time & Success scores as a function of target width for standing and mobile users who received on-target feedback averaged over A/W ratios. 146
Figure 43. Mean steps until selection as a function of target width. 147
Figure 44. Mean time scores for participants that did not receive on-target feedback as a function of distance to target groups and mobility. 147
Figure 45. Success ratios as a function of target width for standing and mobile users (no on-target feedback) 149
Figure 46. Linear Regression lines, mean time scores and standard deviations as a function of Index of Difficulty 150

List of Tables

Table 1. A summary of application areas for Spatial Audio Displays.....49
Table 2. A summary of views on sign divisions and the associated terminology. 55
Table 3. Mean Effective Angle and Standard Deviations for all interaction techniques. 90
Table 4. Experimental Design. The independent variables, orientation update, feedback type and number of display elements and their associated levels. 102
Table 5. Between and within subjects main effects and the interaction between the independent variables in the experiment in time to select measurements. F-values and significance levels are also presented.....105
Table 6. Means and standard deviations for the time measurements in this experiment as a function of display population and feedback cue. O/U refers to whether orientation update was present in the display. The measuremements are in seconds. 107
Table 7. The effect of the independent variables on accuracy scores, F values and significance levels. .108
Table 8. Throughput and effective target width variations between participants. 109
Table 9. Mean Workload Values for the participants that took part in the experiment. In parentheses, standard deviation values are presented. 110
Table 10. Experiment Design: HD stands for Headphones, BC for Bone Conductance, MA for Monaural, A for distance to target and W for target width. The independent variables reproduction and display type are shown together with their associated values and the design choices they had resulted into for the sound positions in the display, their target width and the distances between them. 120
Table 11. Accuracy Success Rates (%) for all Reproduction Techniques, Displays and A/W Ratios. In parentheses standard deviations are given. (HD stands for Headphones, BC for Bone Conductance and MA for Monaural presentation).....124
Table 12. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution for the headphone case.....128
Table 13. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution for the bone-conductance case. 128
Table 14. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution for the monaural case. 129
Table 15. Standard deviations of selections with respect to target sound position. 130
Table 16. Dependent and independent variables used in the experiment and their levels. 140
Table 17. Target Widths (W), Distances to Target (A), and Indices of Difficulty (IDs) and A/W Ratios used in the experiment. 141
Table 18. ANOVA results for participants that received feedback (T is time and S (%) success ratio). M stands for mobility, W for target width and A/W for the ratio of distance to target to target width. 145

Table 19. Goodness of fit comparisons between the linear and logarithmic models. * denotes that correlation was significant at the $p<0.05$ level. R^2_{Lg}, R^2_L, R^2_F stand for the R^2 statistic for McKenzie, Linear and Fitts' models. IP is the index of performance. S stands for standing and M for mobile participants. 150

List of Equations

Equation 1 62
Equation 2 63
Equation 3 63
Equation 4 63
Equation 5 64
Equation 6 64
Equation 7 65
Equation 8 66
Equation 9 67
Equation 10 67
Equation 11 67
Equation 12 68
Equation 13 68
Equation 14 69
Equation 15 69
Equation 16 70
Equation 17 70
Equation 18 70
Equation 19 70
Equation 20 70
Equation 21 72
Equation 22 72
Equation 23 72
Equation 24 72
Equation 25 72
Equation 26 75
Equation 27 76
Equation 28 100
Equation 29 100
Equation 30 100

1 Introduction

The thesis is concerned with interaction design for deictic spatial audio target acquisition. The term target acquisition refers to the process of moving towards the direction of and subsequently selecting a target. The word deictic originates from the Greek word 'δείχνω', which means to point. In effect, the thesis examines a certain type of gesture interaction, the acquisition of a spatial audio target using a pointing gesture. The Oxford English Dictionary [55] defines gesture as

ges-ture, *n.*: a movement of part of the body to express an idea or meaning.

Thus, in the literal sense, pointing communicates the direction of an object or event as this is experienced by an observer. This natural way of communicating has been extensively and successfully used in human computer interaction to accomplish interaction with physical and virtual displays. For interaction with physical systems, we are most often required to move our hands so that we are in contact with a control element and subsequently press a button or move a lever. In personal computing desktop systems, pointing is the basis for all types of interaction supported by modern direct manipulation user interfaces. Any type of manipulation requires users to first point to the desired display element commonly using a virtual pointer controlled by a physical device, such as a mouse.

Direct manipulation interfaces build on a metaphor, a conceptual mapping from the physical reality of the computer to a model world [43]. Such a mapping enables users to interact with familiar entities without having to understand the complex underlying operations that take place inside a computer and are handled transparently by the operating system and the hardware resources. Users interact with the model world as opposed to the computer world, by manipulating the model world elements in a way akin to the way we would interact with such a model world if it had a physical dimension, in this way supporting the metaphor. The most commonly used metaphor is the one of a desktop with entities such as folders and files and application windows coupled with familiar operations like pointing, dragging and dropping and deleting that provide interaction with the model world. The desktop metaphor implemented according to the direct manipulation principles sufficiently describes most current commercial graphical user interfaces for almost all types of applications.

Of critical importance is the provision of immediate feedback with respect to user actions. Feedback closes the loop between human and machine and is there to inform users about the interpretation of their actions by the system. In addition, a major factor that contributes to the success of desktop systems is the utilization of signs, in the form of icons. Icons enhance the clarity and aesthetics of the display, without compromising understandability. Imagine how a desktop interface would look like if only text was used instead of icons. There is psychophysical basis for the use of icons in the display since

users can interact faster with them compared to interacting with text [112]. The design of icons is based on the science of semiotics, the science that studies signs as a mean of communication.

The success of an interface design has to be measured. The concept associated with the success of interface designs is the one of usability. Usability is of multidimensional nature. The varying levels of usability dimensions provide metrics of the success of a design, both in terms of user performance and satisfaction. Usability evaluation enables us to identify problems in interface design and provide solutions that make a display easier to use.

Direct manipulation interfaces rely on pointing to visual items to accomplish interaction. This reliance resulted in the necessity to develop methods to evaluate the usability of pointing techniques. The methodology currently adopted by the human computer interaction community stems from the observations of Paul Fitts [37]. Based on the work of Fitts and other researchers such as Welford [127] and MacKenzie [74], it is possible to characterize human performance in variable pointing tasks, using empirical procedures that include objective measurements of performance and subjective user experience assessment in benchmark tasks.

Although interaction with modern user interfaces is mainly through vision, audition has been found to significantly contribute to the user experience. Audio, when used as a complementary feedback channel, '*essentially succeeds in making the model world of the computer consistent in its visual and auditory aspects and therefore increase users' feelings of direct engagement or mimesis with the model world*' as noted by Gaver [43]. Audio as a complementary feedback channel has been found to effectively increase the usability of widgets. Sonifically enhanced widget design has been extensively studied by Brewster [17]. Such widgets succeed in increasing the usability of graphical widgets by reducing the number of errors users make and the time it takes to recover from them. In particular, in the case of mobile users audio was found to effectively reduce target size, without affecting user performance. When used as a complementary feedback channel, audio has also been found to effectively assist monitoring tasks, such as document download progress and provide context information for users focused in a primary visual task as is found in Chapter 3 . Numerous applications designs exist for this purpose and are described in the literature review.

Audio only eyes free interaction based on pointing has been proposed by Cohen and Ludwig [69] and Edwards [33]. Cohen and Ludwig proposed to take advantage of the directional characteristics of hearing to extend the notion of a window system as this appears in desktop interfaces to the audio domain. They named their concept 'Audio Windows'. In theory, audio windows are an attempt to transfer the direct manipulation principles common in graphical user interfaces to the audio domain. In this way, they provide a general term that encompasses all direct manipulation auditory displays. Edwards [33] proposed that direct manipulation interfaces can be transformed using auditory objects to become usable for blind users. Edwards defined an auditory counterpart for each display element 'auditory objects', defined by their spatial location, a name, an action and a tone. Users were able to get detailed information

on the object's functionality through speech by pressing the mouse button. Edwards approach led to a number of applications that provide audio only presentation of visual displays for blind users.

The thesis tries to examine the applicability of Cohen's views in the context of future human computer interaction. The desktop metaphor has to be seen in the context of technological possibilities of the age of their development. At the time of its development the only display options available for human computer interaction were a screen and a speaker in the computer. Control options were limited to the ones offered by mice and keyboards. A behavioural study of a user interacting with a computer would reveal a sitting person, interacting with virtual icons appearing on a two dimensional screen by clicking the buttons of his control device. It could be claimed that the desktop metaphor was so successful because it was quite suitable for the two dimensional screen people were obliged to use. The virtual pointer controlled by a mouse was such a success because it provided an ergonomic option to efficiently control a virtual pointer on the screen. It took advantage of the fact that computers were placed on a desk to provide good support for the user's hands while they were manipulating the hardware device. Today's technology offers new possibilities that can be used to improve the status of humans interacting with machines.

Nowadays there is considerable work on mobile human computer interaction, pervasive computing, motion tracking devices, wireless internet access and GPS tracking. It is not hard to think therefore that the desktop model will soon become inadequate. Motion tracking devices allow us to capture physical movement. In this sense, it is possible for systems to infer not virtual but physical gestures in 3D space as well as to use physical movement to control virtual pointers. The size of computers is shrinking, already in the market it is possible to buy computing devices which feature location tracking, are internet enabled and fit in a modest sized coat pocket. There is no reason to think of a display as two dimensional anymore. Augmented and virtual reality makes it possible to depict three dimensional objects on top of the real world. Current human computer interaction research is looking for new spaces for display, spaces that can transparently supplement each other in stationary and mobile interaction contexts. Visual based interaction cannot sufficiently cover the new contexts of use [19, 93]. It is interesting to observe that the evolution in screen size is inversely proportional to that of computer size. Although computer size is shrinking visual display size is increasing out of the need to accommodate the ever increasing amount of information we are required to deal with. In this sense it is hard to imagine how the current visual interaction paradigms would fit in contexts different than the common office one. Already small screen devices for mobile use have significant usability problems.

Human computer interaction is now applied in new paradigms and contexts of use. This transformation has to happen both in terms of display and in terms of control. Among other display options such as touch or smell, spatial audio is a promising candidate. Our hearing sense is sophisticated and is capable of working with semiotic information [20, 43, 81]. In principle, most of the information that is presented using vision could be presented using audition. In addition, spatial audio systems that

take advantage of the direction and distance tracking capabilities of our auditory system can enable direct manipulation designs in the audio domain. Most importantly, an audio display is eyes-free, no screen is required and user's vision does not have to be occupied with interaction. Spatial audio displays can therefore provide a feasible, portable way to interact which is extremely suitable for mobile users and eyes free interaction.

Control options will have to comply with the context of use of spatial audio displays, in other words to support mobile and eyes free interaction. One reasonable candidate for such a development is command based control based on speech recognition as in Nomadic Radio by Sawhney and Schmandt [103]. However, such an option is not effective in mobile contexts due to the poor performance of speech recognition technology in low signal to noise ratios and also due to privacy concerns. Gesture control appears to be a more viable option for mobile users due to human ability to perform gestures when mobile without significantly affecting their walking pattern, by taking advantage of our kinaesthetic system. Evidence for this can be found both in the studies by Brewster *et al.* [19] as well in the study by Pirhonnen et al. [93]. These studies demonstrated the potential of audio for interface presentation and gestures as a control option for mobile users.

In particular, pointing based interactions are expected to be a major control option for spatial audio displays. Pointing based interaction succeeds into taking into account the directional information encoded in spatial audio. In this way it provides a good starting point to form the basis for direct manipulation types of design. Pointing can be performed in a number of ways, physically using different parts of the body such as hand, head or virtually by controlling virtual pointers. In this sense, it can serve as a multimodal input technique for the control of spatial audio displays. However, pointing to audio targets has been little studying. Although methods for evaluating the usability of visual target acquisition techniques exist, the same is not the case for spatial audio target acquisition. The thesis aims to fill this gap.

1.1 Motivation & Aims

The thesis builds on the promising concept of Audio Windows [28] and research results that indicate that gesture interaction with spatial audio displays can support mobile human computer interaction. The Audio Windows concept is mostly theoretical and little evaluation has been done. To assess its potential empirical evaluation is necessary due to perceptual problems that occur in virtual spatial audio environments such as localization error and confusions with respect to sound direction [14]. The thesis will investigate the fundamental interaction aspect of 'Audio Window' interfaces, namely pointing based interaction with a spatial audio target. To achieve this, the thesis will seek a framework to evaluate deictic spatial audio target acquisition and identify the factors that affect performance and user satisfaction. The goal is, based on the evaluation results, to propose a target acquisition design that will enable usable interaction.

1.2 Research Questions

The thesis states that to create usable spatial audio displays it is necessary to become able to answer the following questions:

RQ 1	How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?
RQ 2	What are the factors that affect deictic spatial audio target acquisition?
RQ 3	How can we evaluate the usability of deictic spatial audio target acquisition?

Answering these questions will enable the design of deictic target acquisition tasks for spatial audio displays and identify application areas where such interaction is suitable.

1.3 Thesis Walkthrough

The thesis starts with Chapter 2, which examines spatial audio perception in real and virtual environments. This Chapter presents a literature review on the way people perceive the direction and distance of a sound. In this sense, the properties of directional hearing are understood and the problems that can emerge when interacting with spatial audio targets are identified with the goal of guiding investigation on Research Question 1. Chapter 2 also provides background on spatial audio systems which is necessary for understanding the discussion on spatial audio displays in Chapter 3.

In Chapter 3 a review of application designs that use spatial audio is presented. In addition, major design issues that are fundamental for the development and evaluation of direct manipulation interfaces are identified. A review of direct manipulation interfaces is presented and the essential requirements that support the usability of such interfaces are inferred. The literature review also verifies that the research questions the thesis investigates are novel.

Chapter 4 examines in detail the methodology and theoretical background of the evaluation of visual target acquisition tasks or, more specifically, aimed movement tasks. Models of aimed movement to visual targets are presented together with the theoretical perspectives that have been associated with them. A comparison between existing models is provided together with the context in which they are applicable. The chapter also contains a review of the application of the models in human computer interaction and the rationale for the use of the models to provide comparisons between different interaction techniques and prediction of performance in interaction with visual displays. Chapter 4 is purely concerned with visual interaction. Given that no prior research has examined deictic spatial audio target acquisition, research on visually supported aimed movements is used by the thesis to provide insight on answering Research Questions 1 & 3. Chapter 4 concludes the literature review. The rest of the

chapters provide experimental investigations into deictic spatial audio target acquisition in the horizontal plane.

Chapter 5 presents an experimental evaluation of the feasibility of spatial audio target acquisition in an exocentric display, the importance of feedback marking the target area and the effectiveness of three acquisition gestures. This chapter is mostly related to Research Question 1 and 2. Feedback is found to be crucial for the effective interaction and the effect of pointing gesture on spatial audio target acquisition is confirmed. Pointing using the hand is found to be promising enough compared to nodding using the head and selecting using a stylus on a touch tablet and is taken further.

Chapter 6 compares interaction in egocentric and exocentric displays with or without a number of feedback cues in the presence of distracters and their effect on interaction effectiveness and efficiency. An evaluation method inspired by visual target acquisition studies is proposed to assess the effect of the different feedback cues. Interaction in egocentric designs is found to be fast but not accurate while interaction in exocentric designs using the specific selection task is found to be slow but accurate. Feedback marked spatial audio elements are found to compensate for time and accuracy deficiencies in both cases. The investigations in this Chapter are related to all Research Questions and provide the necessary data so that a step forward in the design of spatial audio target acquisition can be taken.

In Chapter 7, the results of the experiments in Chapter 6 are used to create a spatial audio acquisition task that is evaluated under three reproduction options and two display designs in the presence of distracting sounds. The results indicate that Fitts' law could be used to model spatial audio target acquisition in the frontal horizontal plane when the display is presented using headphones and is a research direction that should be taken further. Again all Research Questions are addressed in this Chapter. An evaluation approach inspired by visual target acquisition is found to be promising and effects of spectral and binaural cue deprivation, target width and the ratio of distance to target over target width are found.

Chapter 8 presents an evaluation design that can be used to evaluate spatial audio target acquisition using pointing gestures that can be applied in the case of mobile participants. The design is applied on the selection task presented in Chapter 7. The effects of target width and the ratio of distance to target to target width on spatial audio target acquisition are examined for the case of standing and mobile participants. Fitts' law is found to account for spatial audio target acquisition. The Chapter again investigates all Research Questions. With respect to Research Question 1 it verifies the necessity of feedback marked audio areas to support spatial audio target acquisition and identifies effects of mobility on spatial audio target acquisition related to Research Question 2. The evaluation method proposed is found to successfully examine a multitude of aspects necessary for the successful design of the interface and therefore contributes to Research Question 3.

Finally, Chapter 9 summarizes the contributions and findings of the thesis as well as presents the limitations of the research and future work that could undertaken to resolve them.

2 Spatial Sound Perception in Real & Virtual Environments

2.1 Introduction

The goal of this chapter is to examine the literature in spatial sound perception in real and virtual environments and to provide background on virtual spatial audio systems and their limitations. The reason for investigating perception of sound location in real environments is because it sets a benchmark for the fidelity that can be achieved in virtual environments.

Spatial sound perception in real environments is examined in psychology literature mostly by empirical studies. Similar methods are used to evaluate virtual spatial audio systems. Existing literature provides the physical parameters that affect spatial sound perception as well as the limitations associated with sound localization.

Virtual audio systems employ binaural technology together with signal processing techniques to create the impression of directional hearing. Virtual audio environments have been evaluated in detail from a psychoacoustical point of view and their fidelity has been compared to the real world. Such evaluations and comparisons are usually done on grounds of localization accuracy. In general, perception of sound location in contemporary virtual environments is not as accurate as in the real world. This is mainly attributed to technological limitations.

Due to the fact that the thesis deals with the acquisition of spatial audio targets an understanding of spatial sound perception is necessary. Estimates of the ambiguity inherent in the location of sound events in virtual environments are of particular interest, the goal being to assess the effect of localization problems on interaction and to recommend design solutions.

2.2 Theory of Sound Localization

Before going into the details of the theory of sound localization, it is useful to distinguish between physical events in the real world and perceived events created based on information available through our senses. In this sense, the location sound event has two dimensions depending on the frame of reference: a physical, objective one and a perceptual, subjective one. Subjective estimates of the location of a sound event are obtained based on temporal and spectral information arriving in our auditory system. Theories of sound localization are trying to establish how this process is performed. In this context, sound localization is defined by Blauert [14] as '… the law or rule by which the location of an auditory event (e.g. its direction or distance) is related to a specific attribute or attributes of a sound event or of another event that is somehow correlated with the specific event'. Studies of sound localization have revealed that there is a certain ambiguity in the location of sound events. The term localization error is used to refer to

the uncertainty in sound position estimations, as these is revealed in the responses of people asked to judge the direction of a sound event in a certain way. The term Minimum Audible Angle (MAA) or localization blur refers to the smallest perceptible change in the direction of a sound. Two attributes define our impression of the location of a sound: its direction and its distance. In the context of the thesis, localization error is more relevant than localization blur since it represents the error in listener's judgements rather than the resolution of the auditory system, in a real or a virtual auditory environment.

The position of a physical or perceived sound event is defined relative to the head of the listener using three variables distance, azimuth and elevation. They are illustrated in Figure 1.

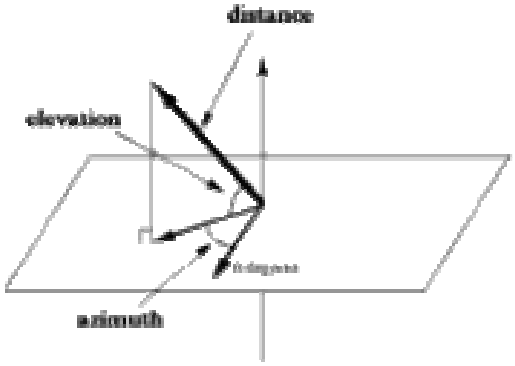


Figure 1.An illustration of azimuth, elevation and distance, the variables that are used to define the position of a sound source relative to a listener's head. 0° of azimuth corresponds roughly to the direction of a user's nose. The point of reference is assumed to be in the centre of the head, adapted from [61].

In addition, the space around the head of a listener is partitioned according to Figure 2 for ease of reference. This is done to provide a common way to refer to certain areas around our heads that are interesting from a psychoacoustical point of view. To give an example the median plane defines an area that is symmetrical with respect to the ears of a listener and the sound pressure variations at the ears of a listener for sounds positioned in the median plane are quite similar. For most locations in the horizontal plane, the sound pressure at the ears of a listener differs substantially due to the interference of the head, a fact that is considered beneficial in determining the location of the sound. This is the reason that all spatial audio display designs presented in Section 3 used sounds positioned in the horizontal plane.

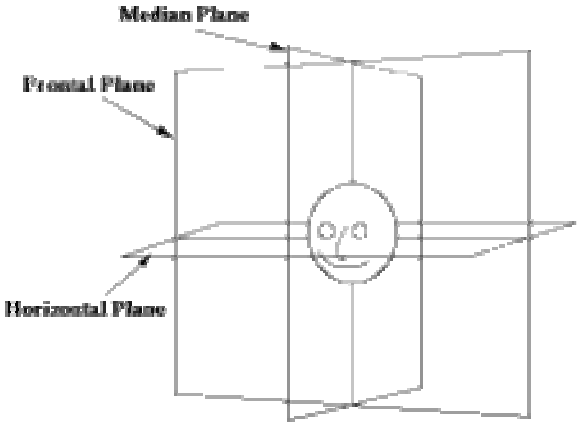


Figure 2. Illustration of the median, frontal and horizontal frames. These areas are commonly used to refer to the location of sounds, relative to the head of a person due to the fact that they result in certain physical properties for the signal that reaches the ear, adapted from [61].

Auditory localization theories use concepts from The Duplex Theory by Lord Raleigh [98]. The Duplex Theory explained human sound localization ability based on the differences of sound pressure that arrive at the two ears. These differences are called interaural or binaural differences. They are mainly interaural level differences, interaural time differences in the frequency range of up to 1.5 kHz and interaural time differences between the envelopes of the signals. These differences occur naturally because of the presence of the head. For example, a sound emitting for the left side, effectively reaches the left ear faster than the right ear and the sound pressure level on the left ear is higher than the one on the right due to certain frequencies being reflected by the head. This phenomenon is also called head shadowing [85].

Interaural level differences are particularly prominent in high frequencies (over 1.5 kHz) where the wavelength can be considered to be small in comparison to the head. At such situations level differences can be in the order of 35dB. On the other hand they are negligible for frequencies below 500 Hz. Interaural time differences range from 0 (for a sound straight ahead) to about 0.69 ms for a sound at the sides. Interaural time differences are particularly important for the localization of low frequency sounds where it is suggested that they dominate perception. At high frequencies, many cycles of phase difference between the ears can appear and for this reason this cue is ambiguous.

Interaural differences can sufficiently explain sound localization in the frontal horizontal plane. However, the theory cannot account for certain areas where the information arriving at the ear cannot be uniquely associated with a single position in space. These areas include the median plane and the points on what has been named, the ‘cone of confusion’ [85]. The cone or cones of confusion are loci of constant interaural time and intensity differences. In Figure 3, it can be seen that points on the cone will result in the same interaural time and intensity differences.

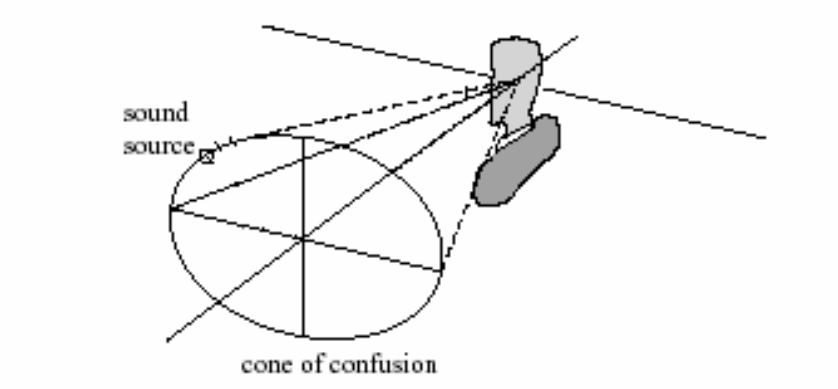


Figure 3. An illustration of the cone of confusion. Sounds located around the cone result in the same interaural time and intensity differences, and it is thought that this the reason that their position cannot be reliably identified in the absence of pinnae cues, adapted from [97].

The apparent inability to recognize the position of sounds in certain areas is not only a limitation of the model but in certain cases it is also a perceptual limitation. This limitation is manifested in the phenomenon of confusions. Confusions are most commonly front-back but up-down confusions have also been observed. In such situations, a sound that is located in front (or above) a listener appears as coming from behind (or below). For broadband sounds played through loudspeakers confusions may appear at a rate of up to 10% [83]. This phenomenon is particularly strong in the median plane. Blauert [14] mentioned that in the median plane a signal that contains energy in specific frequency bands is likely to be judged as coming from the front or back irrespective of where the sound is actually located. This phenomenon is really extreme for narrow band signals having bandwidth of less than 2/3 of an octave [14, 83]. In this case, localization depends not on the location of the sound but only on its spectral content. As more recent research [83, 87] indicates spectral cues play an important role in localization as long as elevation is concerned. These cues result primarily as an effect of our outer ears. As the sound reaches the outer ear it undergoes certain distortions due to reflections. These distortions are direction dependent and result in amplification or attenuation of certain frequency components. It is believed that the auditory system associates distortions to certain directions therefore: they play an important role in localization for non-zero elevation values. For this reason they form part of theories of spatial hearing.

In conclusion, the auditory system utilises binaural differences in combination with spectral cues to estimate the direction of a sound event. The perceived direction of a sound can be predicted to a certain extent based on interaural differences and its spectral content. However, due to the empirical nature of most of the studies of spatial hearing, it is relatively difficult to form a broad picture of how exactly the auditory system estimates the location of sound events. Most of the studies study localization in confined areas of space and with certain usually narrowband stimuli. In addition, studies try to simulate what is

called the free field that is environments where reflections are eliminated. In this sense, the results are valid in their own context and indicative of the abilities of our hearing system however they might appear as conflicting with studies in other experimental contexts and might not be directly applicable in real world situations. The next section provides details and estimates on localization accuracy in a number of situations that are relevant to the development of the thesis.

2.3 Spatial Sound Perception in Real Environments

According to the previous section, the three most prominent cues from psychology studies that can be used to display information using spatial audio are direction, distance and motion of sound events. Perception of direction has been widely studied and has been shown to be consistent for people with normal hearing. In addition, to a certain extent it can be explained theoretically. In this sense, existing literature can help a designer understand the limitations and design directional interactions with sound events. Perception of distance has not been studied in detail. Some results exist and overall indicate that judgements on distance depend on factors that sometimes cannot be controlled by design such as familiarity with the sounds [84]. However, it is found that there is some consistency in judgements and the possibility of using distance cues in the display should not be completely abandoned. Detection of motion essentially combines direction and distance perception. In addition, moving displays have the advantage that they are attention grabbing. The literature indicates that motion can be a useful design tool for the display of appropriate information.

2.3.1 Notes on the Perception of Direction

Localization error has been measured in setups involving real sound sources. In the past, most studies aimed at estimating localization error in the horizontal median planes due to technological limitations mostly associated with loudspeaker positioning. However, more recent studies employ multiple loudspeakers arrangements and in this sense provide a better picture on how localization error varies in various directions around the human body.

Localization error primarily depends on the position and the spectral content of the target sound. When judging the azimuthal position of a sound event, localization ability is dominated by the interaural cues [14, 84]. Neither the spectrum of the sound nor the outer ear influence localization ability significantly. However, for sounds at the sides as well elevated sounds and sounds to the rear of a listener, spectral cues due to the distortion effect of the outer ear (pinnae) becomes important and in some cases dominant. This is due to the fact that interaural cues become ambiguous in the loci of cones of confusion. The effect of pinnae is quite important for high frequencies, due to the fact that at longer wavelengths the interaction with pinnae is negligible. Middlebrooks [84] puts the threshold after which the effect of the outer ear becomes important at 4 kHz while Moore [85] at 6 kHz and more.

The region of most precise spatial hearing lies in the forward direction of the horizontal plane [14]. Figure 4 provides an illustration of how localization error varies for broadband sounds positioned in the horizontal plane. It is evident that there is substantial variation and a maximum value of ±10° is observed at the sides of the listener. Results are verified in a more recent study [84] where smallest errors for azimuth are in the order of 2° and increase to 20° for certain areas to the rear of the listener.

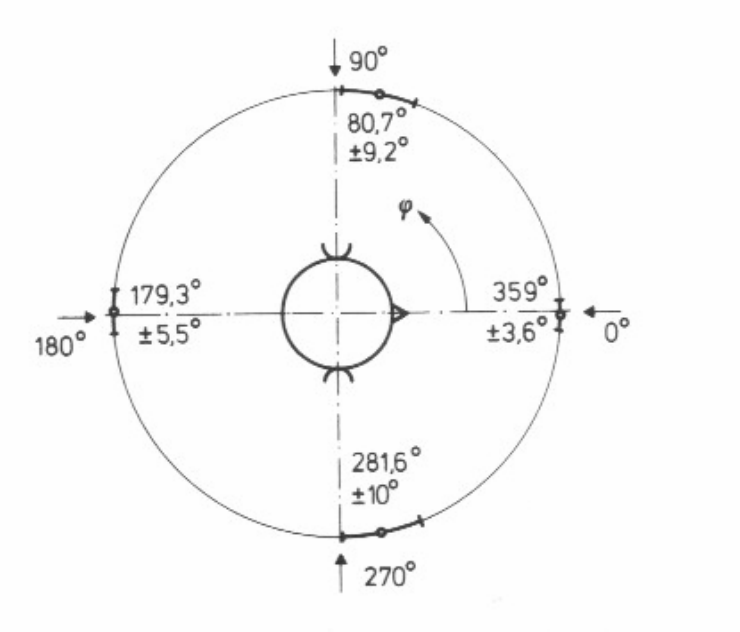


Figure 4. Localization error at a number of angles, adapted from [14]. Mean perceived direction as well as its deviation are illustrated for sounds originating at the direction of the arrow.

Apart from localization error, another way to estimate localization error is through estimating the minimum audible angles. Minimum audible angles (MAA) correspond to the smallest detectable change in angular position. This is essentially different than localization error in the sense that listeners are asked to discriminate whether two sounds come from the same position or not as opposed to judging the direction of a single sound. In a study by Mills with sinusoidal stimuli (referred in [85] pp. 218), MAA was smallest (about 1°) for stimuli directly in front of the listener for frequencies up to 1kHz. Performance worsens around 1500-1800 Hz. When the reference direction is moved away from 0° in azimuth MAA can grow up to 7° for low frequencies, however for frequencies above 1.5 kHz it may grow so much that for certain directions it cannot be estimated. Data are presented in Figure 5.

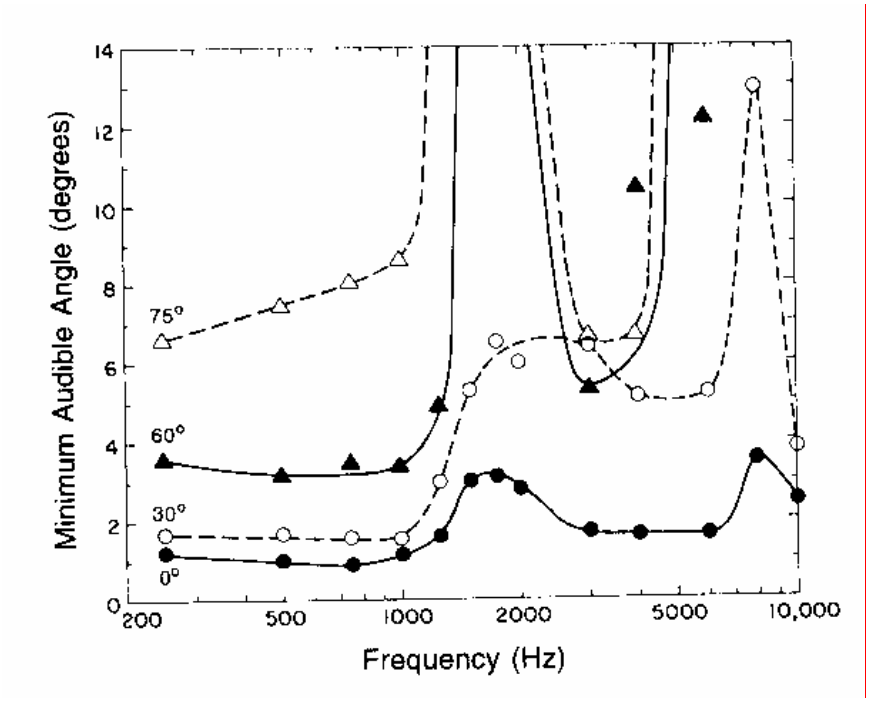


Figure 5. The minimum audible angle (MAA) for sinusoidal signals, plotted as a function of frequency, each curve shows results for a different reference direction. As can be seen for certain directions of a sound, the minimum audible angles can vary a lot as a function of frequency, adapted from [85].

Measurements of sound localization were also performed by Oldfield and Parker [89, 90]. The authors found that in the case of absent pinnae cues increased substantially elevation error and the number of front-back reversals. The authors also found that when pointing is used by subjects to indicate sound source position substantial motor error is found for sounds at the back of the users, which is not found when other methods of declaring sound position are used.

Localization ability is degraded for areas influenced by the cone of confusion phenomenon (see Figure 3). This is observed in higher localization error values and front-back and up-down confusions. The region of most inaccurate perception is in the median plane. For broadband sources, localization error in the median plane varies from ±9° at 0° elevation, ±10° at 36° elevation, ±35° at 90° elevation and ±15° at 144° elevation [14, 84]. For elevated sources elsewhere, the azimuthal component of localization error is usually close to the one observed in the horizontal plane however, the total error is higher mainly due to elevation.

Localization error depends to a certain extent on the frequency content of the sound. This dependence is largely associated with the effect of pinnae (the outer ear) on sound waves reaching the ear. We are better at estimating the location of broadband sounds with substantial spectral variation over time [85]. To give an example, for sounds displaced forward in the horizontal plane, there is experimental

evidence that localization error varies from a minimum value of about 1° for impulses (clicks) to a maximum of 4.4° for sinusoids. It is believed that localization in the horizontal plane is relatively invariant to the spectral content of a sound source as binaural cues dominate perception. In the median plane, however, the spectral content of the sound source is important. This, in effect, influences the perception of direction for sounds located on top and to the back of a listener. Again, broadband signals result in smaller localization error values. The concept of directional bands has been developed by Blauert to account for perception of sound direction in the median plane in relation to the frequency content of a signal. Figure 6 provides an illustration of the concept of the directional bands as these have been estimated experimentally.

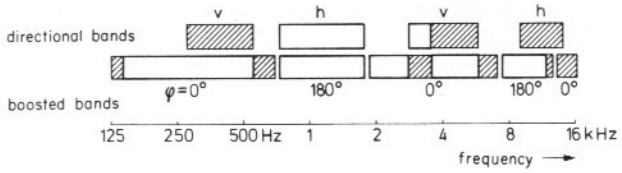


Figure 6 Illustration of the concept of directional bands, adapted from [14]. H stands for behind, v for forwards. Sounds containing frequency components inside the bands are perceived as coming from backwards or forward, irrespective of their actual direction.

Relevant to the concept of directional bands is the concept of boosted bands. Boosted bands correspond to frequency regions that are boosted by the outer ear. These regions can be identified in certain peaks and dips of the outer ear transfer function, that are direction dependent. Such frequency bands however, can only be coarsely thought to be common among people. The phenomenon might not be possible to describe in such a consistent way across people due to individual differences. In the cases where narrowband noise on a band that corresponds to a certain direction dependent peak of the transfer function of the outer ear is played in isolation, localization responses correspond to the particular direction [14, 84].

However, such observations can only be associated with certainty with bandlimited signals. When the signal is broadband, the relative salience of the frequency bands might not be the same. An issue that is debatable is to what extent spectral cues are binaural or monoaural. Monaural cues have been found sufficient for localization in the vertical dimension (elevation) and in limited cases of people with hearing impairment in one ear, moderately successful for localization in azimuth. However, there is no confirmed body of research on the matter as yet.

Head movements are also thought to assist in the perception of direction of sound events lying on cones of confusion and in the median plane. It has been argued that head movements provide cues that help disambiguate sound position when confusions occur. However, recent studies show that although head movements help to reduce confusions they are not eliminated [21]. The possibility of head

movements is also dependent on the duration of the stimulus. For short stimuli (less than 300ms) head movements are not observed. In most of the studies people were able to localize elevated sounds while their heads were kept still. So head movements can be thought to act complementary to spectral cues. Familiarity with the sound source is also helps in alleviating confusions [85].

Finally, it is worth noting the 'precedence effect'. The effect is relevant in situations where reflections are present in the environment, as is quite commonly the case in enclosed spaces. In such situations, perception of sound direction is dominated by the direct sound that reaches the ear. Reflections, as long as they are close enough in time to be perceptually grouped with the direct sound event and not be perceived as echoes, are not influencing perception of direction. This fact explains our ability to localize sounds in enclosed spaces.

2.3.2 Notes on the perception of distance

Apart from sound direction, distance to sound events is a potentially promising design tool. From a perceptual point of view it has been found to depend primarily on the sound pressure level at the position of the listener. In general, judgements of absolute distance are in better agreement with actual distance when listeners are familiar with the sound source. Such familiarity may be achieved through training, or through everyday experience.

An acoustic analysis is useful in assisting the understanding of the perception of distance. This is because the influence of the head on the acoustic field varies with the distance to the sound source. For distances less than 3m the listener can be considered to lie in the near field of the sound source. In such cases, the curvatures of the wavefront arriving at the head cannot be neglected. In this sense, the spectrum of the ear input signals cannot be considered uniform and it changes at different positions of the listener. For distances between 3m and 15m the only thing that depends on distance is the sound pressure level. The influence of the head on the sound spectrum is minimal. At distances of more than 15m sound pressure level is still important but attenuation of the high frequency components of the sound is also observed, due the sound wave propagating through air. Additional factors that are considered by the auditory system are reflections and reverberation. Normal listening experience involves listening to sounds in enclosed spaces. In these situations, reflections and room reverberation come into play.

In medium distances between 3m and 15m and in the free field, a doubling of distance results acoustically in a 6 dB reduction of sound pressure level. However, it has been shown by von Bákesy as well as Laws referenced in [14], that a doubling of the perceived distance is achieved by reducing sound pressure level by 20dB. This implies that when judging distance based on sound pressure level the perceived distance increases less rapidly than the actual distance of the sound source itself. Von Bákesy referenced in [14], hypothesized because of this reason, the auditory space is of a limited total extent and there is an outer limit to the distance of auditory events, called the 'auditory horizon'.

Figure 7 illustrates how perceived distance varies as a function of actual distance from sound for three levels of speech: whispering, normal speech and calling out loudly. It can be shown that the aforementioned phenomenon appeared only for whispering. For normal speech there was a certain correspondence between perceived and actual distance to sound event and for calling out loudly the distance to sound event was overestimated. The results are somewhat surprising given that distance for calling out loudly was estimated to lie in greater distance that normal level despite the higher sound pressure level. This fact stresses the importance of familiarity with the signal and its context. Evidently calling out loudly is associated by everyday experience with talking to someone at a larger distance. a fact that biases our estimations of distance.

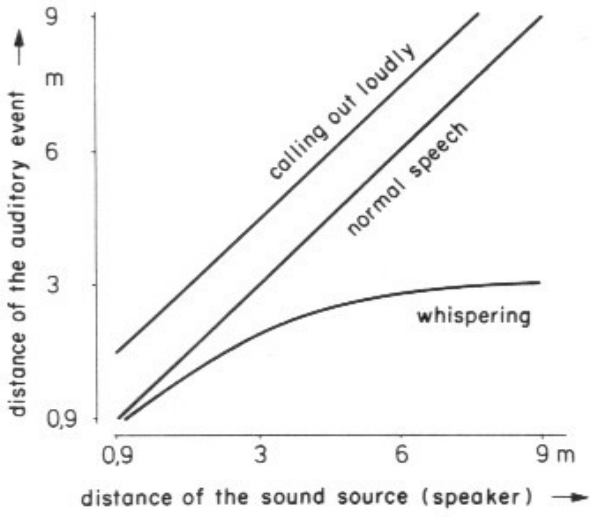


Figure 7, Perception of distance vs. the actual distance from the source, adapted from [14]. If an one to one relationship between actual and perceived distance the result would be a straight line with a slope of 45°.

Ashmead *et al*. [6] have shown that listeners can discriminate differences in the distance of sound sources as well as could be expected from the optimal use of intensity cues. In addition, in [5], it was shown that the change in intensity that occurs when a listener walks towards a sound source can provide an absolute cue to distance. This cue was effective even when the overall intensity of the source was varied randomly from trial to trial. This result is quite important since it shows that absolute judgements of distance that otherwise are difficult to achieve unless high familiarity with sounds exists, can be achieved through moving towards the sound.

For distances greater than 15m, attenuation of high frequencies due to propagation becomes significant. This cue works together with further attenuation of sound pressure level. It has been found

that this cue is used by the auditory system in judging relative distance e.g. far, rather than absolute distance.

For sounds in distances less than 3m, apart from sound pressure level changes spectral changes due to the interaction with the head come into play. It is believed [14] that the auditory system evaluates distortion of the spectrum of the ear input signals besides evaluating sound pressure level. At distances higher then 25 cm the influence of the spectral distortions is small. It has not yet been determined which spectral attributes are evaluated.

In most everyday situations listening is performed in rooms. In this case a certain factor that does not appear when considering the free field is reflections. It has been found that the ratio of direct to reflected sound and the time delay between the direct and the reflected sound provides cues to distance [85].

Another issue that is associated with distance hearing is what has been called 'inside-the-head locatedness' (IHL) [14]. This term refers to the phenomenon where a sound source is perceived to be located 'inside' the head of the listener. This a common phenomenon especially related to headphone reproduction. The reason for IHL is not particularly clear. However, Laws referenced in [14] was able to verify experimentally that IHL is related to the difference of the signal when presented by loudspeakers versus when presented by headphones. When sound is reproduced by headphones, sound waves are emitted on the axis of the ears and very close to the entrances of the ear canals. This can lead to linear distortions of the ear input signals. Moreover, factors such as attenuation of the signal caused by the air medium are not present when sound is reproduced by headphones. Laws, was able to avoid this effect by equalizing the signal from headphones to become similar to ear input signals when sound was reproduced by loudspeakers. According to this result free field equalization of headphones seems to be necessary to avoid the unpleasant IHL phenomenon.

2.3.3 Cross Modal Effects

It has been argued that cross modal effects, occurring because of visual information can influence the perception of the location of a sound event [110]. Perceived body position and perceived body motion bias auditory spatial perception. When vestibular cues give rise to illusory perceptions, exocentric auditory localization judgements are biased. The term exocentric refers to judgements based in a frame of reference not residing on the body of the listener. In such cases, the responses of participants are affected by their perceived body positions.

Auditory spatial judgements can also be biased by visual or proprioceptive spatial information. In such cases, the auditory sources are displaced towards the location of the visual or proprioceptive stimulus. When visual and acoustic spatial cues are in conflict, the bias is so strong that listeners sometimes perceive the auditory stimulus as arising from the same point as the visual stimulus. This is known as the ventriloquism effect as mentioned in [110]. Visual cues bias auditory perception to a greater

extent than auditory cues bias visual perception. Auditory localization judgements are typically displaced by about 50%-80% of the difference between the visual and the audio stimulus. Audio cues bias visual perception to a degree of about 6%. The same situation applies to bias by proprioceptive stimuli.

The above values are however dependent on experimental conditions and represent extreme situations. When people know that differences might exist between the competing modalities the effects are notably smaller. However, it should be noted that visual aids, such as reference frames, help auditory localisation judgements and they assist spatial information processing.

2.3.4 Perception of Auditory Motion

Auditory motion is an interesting cue from a design point of view, due to the fact that it is attention grabbing. Moving sounds might have a variety of applications in human computer interaction, since they provide the opportunity to sonify moving display elements that can be used in games and other interaction contexts. There is debate in the literature on whether auditory movement detection is done using the same subsystem that is used for the detection of direction and distance. That is, researchers question whether the auditory system has developed specialized neural elements for the detection of motion. Grantham [49] hypothesized that movement detection is done based on perceptual estimators of the position difference between the trajectory onset and termination and the elapsed time. However, Perrott *et al.* [91] showed that listeners could discriminate between targets that travel the same distance in the same time and in the same trajectory, but use different acceleration and deceleration patterns, implying that either listeners sample the trajectory in more than one point or more cues are used for the determination of auditory motion.

Lutfi and Wang [72] provide an analysis of the relative reliance of the acoustic cues that are used to discriminate auditory motion as well as thresholds for the discrimination of two moving sound sources. They have found that at moderate velocity levels of 10m/sec, intensity and interaural time differences are preferred for the discrimination of displacement, while the Doppler effect expressed as the rate of change in frequency, is preferred for the discrimination of velocity and acceleration. At high velocities, in the order of 50m/sec the Doppler effects prevail in the discrimination of auditory motion. Thresholds for the discrimination of two auditory targets were expressed as an equivalent change in incident angle over the total duration of the signal. They were 15° for displacement discrimination, 11° for velocity discrimination and 8° for acceleration discrimination.

It is, however, the case that localization of moving sources is not as accurate as the one of static sources. In a study by Grohn *et al.* [50] it was found that localization error for moving sources was at least double the one for static ones.

2.4 Spatial Audio Reproduction in Virtual Environments

Spatial audio reproduction in virtual environments is achieved using principally two methods. The first of these methods is called ambisonics. Ambisonics [78] or surround sound techniques attempt to reconstruct the acoustic field as this would be if the sounds were physically present in the environment, by using loudspeaker arrangements. A number of different methodologies have been developed for rendering the sound field, see for example [12, 86, 88]. Different loudspeaker arrangements have been proposed in the literature, ranging from four, to loudspeakers arrays that use large numbers of loudspeakers in geometric arrangements such as cubic or spherical. This technique provides good spatial audio quality; however, its main disadvantage is that the spatial impression effect only appears in a certain area of the space enclosed by the loudspeakers, which is usually called the 'sweet spot'. In addition, this technology is not mobile due to relying on loudspeaker reproduction; therefore users have to be confined in space. For this reason, its application is mostly in spaces such as living rooms, theatres and cinemas. These properties make this technique unsuitable in the scope of the thesis and for this reason it is not reviewed further.

The second option for spatial audio production is HRTF (Head Related Transfer Function) filtering [14, 44, 85]. HRTF filtering is based on the assumption that humans will tend to perceive the same auditory event when presented with the same ear input signals as the ones occurring in the real world. HRTFs are measured empirically and capture the properties of the acoustic path to the inner ear, including the effects of the outer ears (pinnae). When HRTFs are applied to a monophonic sound signal they result in a binaural signal that is perceived as emitting from a given direction in space. HRTF filtering can be implemented in real time and can thus provide a portable way to produce spatial audio. HRTF reproduction can be performed either using two loudspeakers as is examined in detail by Gardner [41] as well as using headphones as for example in Wenzel *et al.* [128]. In the former case, it is necessary to perform cross-talk cancellation to avoid interference between the left and right ear input signals. Loudspeaker reproduction is not relevant to the thesis, due to the requirement that interaction could be performed in a mobile setting. Loudspeaker reproduction would require that the user is confined in space and therefore cannot support mobile interactions. For this reason only the case of headphone reproduction is considered in this chapter.

2.4.1 Details on HRTF Estimation and Implementation

The invention of HRTF filtering is largely based on psychoacoustics research. As discussed in Section 2.2, the auditory system localizes sounds using interaural differences for azimuth and spectral cues for elevation. The spectral cues are mostly due to the effect of the outer ear. All cues might be combined in a transfer function that captures the properties of the path from the sound source to the inner ear.

Estimation of the transfer function can done in the context of a psychoacoustic procedure [14, 42]. Small microphones are inserted in the ear cavity to record the signal that arrives to the ear. In the context of HRTF measurements, the external ear is assumed to be a linear time invariant filter and thus its frequency response at different angles can be estimated by dividing the power spectrum of the signal at the beginning of the ear tube by the power spectrum of the original signal played. Such measurements are done with sounds emitting from loudspeakers from a number of locations that sample the area around the listener's head. A number of such measurements is averaged to reduce the noise of the measurements. In this way, a database of angular frequency responses is created that can be subsequently used to position a sound signal arbitrarily. For unknown positions the frequency response is estimated by interpolating the closest known frequency responses.

Due to the fact that the process of estimating HRTFs for individuals is laborious, researchers tried to use HRTFs from mannequins built in a way that approximates an average human outer ear, head and torso or from a single 'representative' listener. These HRTFs are called generalized or non-individualized [14, 128]. The advantage of using generalized HRTFs is that they provide a general way to create 3D audio, without the need for individual measurements. The disadvantage is 3D audio fidelity degradation. The effect is discussed in Section 2.4.2.

HRTF measurements are assumed to be free field measurements. They are usually performed in an anechoic chamber that eliminates reflections. However, it is argued that if sounds are synthesized in the free field, localization is not as accurate and the result effect is unnatural compared to sound rendering that takes into account reflections from a natural reproduction space. For this reason it is suggested that further treatment of the sounds should be performed to imitate reflections from the room. However, this feature is not always available in commercial implementations.

HRTF filtering is computationally intensive, and in order to achieve a fast implementation a digital signal processing technique called sectioned convolution has to be performed to the input signal. The filtering is done in the frequency domain using one of the digital signal processing methods available for this purpose [61]. By using HRTF filtering, sound can be positioned at an arbitrary point around the user. However, sound localization is not as accurate as in real life and the possibility of having more than one sound can furthermore error the perceived sound position [50, 68].

Finally, for simulation of virtual 3D worlds it is important to provide real time update of the sound positions relative to the direction and the distance to the listener. This is usually achieved by using tracking devices placed on the head of listeners. In this way, the orientation and distance of listeners to the sources in the 3D world can be estimated and this information can be used to update sound positions relative to the listener. However, this means that the sounds have to be filtered frequently to appear as coming from their new positions. This process requires significant computational power and limits the application of this technique in low computational power devices. In addition, for convincing realization the update rate has to be very fast (less than 10 ms), a timing requirement that is not possible in most

current operating systems. In effect, most people notice the process, however most can adapt to it and take advantage of head movements to disambiguate confusions with respect to sound position. The effects of lag are also discussed in detail in Section 2.4.2.

2.4.2 Sound Localization in systems using HRTF filtering

HRTF filtering provides relatively good spatialization, however the localization error and the number of reversals in an HRTF based system is larger than in the real world, especially when using generalized HRTF functions. Wightman and Kistler [131] compared localization using loudspeakers to localization using individualized HRTF functions. They found that judgements of the azimuth of sound position when using real sounds from loudspeakers and when using headphone presented individualized HRTFs filtered signals were correlated at 98% with actual sound positions in both cases. With respect to elevation their figures were 90% and 82%. Confusions were at a rate of about 6% when using loudspeakers and 11% in the virtual case. Mean absolute error of judgement ranged from 16° to 30°, when trained participants had to verbally report sound direction. Their results indicate that HRTF filtering using individualized transfer functions can successfully create a spatial audio impression with localization errors comparable to the real world sound localization. However, it was found that confusions are more likely in a virtual than a real environment. Another important observation was that there were large individual differences between participants. Some were significantly more able to localize sounds than the rest both in free field and in the virtual environment. In addition, it should be noted that the participants in Wightman & Kistler's experiment were experienced in sound localization judgements. Sound localization estimations were spoken out by the participants in degrees.

Another study that evaluated the fidelity of three dimensional sound reproduction using a virtual auditory display was the one by Langengijk and Bronkhorst [65]. This study also compared sound localization in the free field to sound localization using HRTFs and headphones, however the comparison is not made in terms of absolute localization judgements, rather on whether listeners could discriminate real from virtual sounds. The experimenters used three experimental methods to test the discrimination of real and virtual sources. The first one was a yes-no task where listeners had to indicate whether they were listening to the real or the virtual source, the second was a 2 alternative forced choice method and the third was an oddball design with four intervals, where listeners had to identify in which quadrant around them a sound was located. Experimenters used individualized HRTF functions and in addition an acoustically open transducer placed in front of the ear of the listener, rather than headphones.

Figure 8. The acoustically open transducer used by Langendijk and Bronkhorst. The trnsducer is positioned at a distance to the ear to allow for a microphone to be placed in the ear.

Sounds were played either from a loudspeaker positioned in one out of six possible locations or were HRTF filtered and played from the transducer shown in Figure 8. Listeners were asked whether they could discriminate between the real and the HRTF filtered sound stimuli. The results indicated that for all methods listeners could not discriminate between the real and virtual sound above chance levels.

This experiment also investigated whether interpolation of HRTF functions affected localization performance. Interpolation is used to calculate an HRTF set for an unknown position from known HRTF sets that correspond to neighbouring positions. The experimenters examined the effect of direction of interpolation, the interpolation interval and stimuli variability on the discrimination tasks between real and virtual sounds. They found that interpolation direction produced a significant effect with vertical and horizontal interpolation rating better than diagonal. In addition, consistency of the stimuli was found to improve listeners' ability to discriminate between the real and virtual sounds. When a flat noise spectrum was used, listeners were better at discriminating sounds compared to the case when the energy contained at different frequency bands of the stimuli was varied randomly between trials. This indicates that listeners use timbral (spectral) cues that can be uniquely associated with certain sound directions when making a localization judgement. Finally, the listeners were not able to perceive the difference between interpolated and real measured HRTFs when the interpolation interval was less than 6°. This was correlated to the magnitude of the error in HRTF estimation that is introduced by the interpolation procedure. The researchers state that acoustical differences between 1.5 and 2.5 dB per 1/3 octave result

in timbre differences that are not associated with localization by listeners. If however the error exceeds 2.5 dB then it can affect the judgement of sound location.

Bronkhorst [21], evaluated localization for real and virtual sound sources using individualized HRTF functions. The sound stimulus was a harmonic signal with a fundamental frequency of 250 Hz and between 16 and 60 evenly spaced harmonics. The listeners that participated in the experiment performed localization judgements using two different tasks. In the first the so called 'head pointing' task, the stimulus was played repeatedly until a localization response was obtained from the participants. Participants had to turn their heads until they were facing the target sound. The second task used only short stimuli and did not involve head movements. In this task the participants of the experiment had to indicate the quadrant of the horizontal plane from which the sound originated as well as whether the sound was above or below them. These tasks were performed for both real and virtual sound sources. Two HRTF filters were used one that contained 256 points and one with 1024 points. The longer filter effectively contains a larger portion of the impulse response. 13 positions in the azimuth and 7 in the elevation axis were used for stimulus presentation. In the pointing task only positions in front of the listeners were used. In addition, the bandwidth of the stimulus was varied, stimuli contained either 28 or 60 harmonic components for the real sounds and 16, 28, 40 or 60 components for the virtual sounds. For the head pointing task mean absolute errors were shown to reduce as the frequency range of the stimuli was increased. For virtual sources mean absolute error was about 20° for a cut-off frequency of 4 kHz and was reduced to 14° for a cut-off frequency of 16 kHz. For real sources, the variation was between 14° for a cut-off frequency of 8 kHz to slightly less than 10° for a cut-off frequency of 18 kHz. Localization error was found to be significantly greater for virtual compared to real sounds. This variability was mostly due to elevation errors, errors in azimuth judgements were comparable between real and virtual sources. Response time was also found to reduce based on source frequency, from 6.1s to 4.8s on average. It was not however, affected by source type and source position. The confusion task was not targeted towards accuracy but rather towards confusion percentages with respect to sound direction. The rate of confusions was found to be significantly higher for virtual compared to real sound sources and decreased significantly as signal cut-off frequency increased. However, the frequency dependence was verified only for real and not for virtual sources. Filter size was not found to significantly influence the results. The results of this study indicate that significant localization cues are contained in the frequency region above 7 kHz that were not appropriately simulated by the HRTF filtering technique employed.

Wenzel *et al*. [128] investigated localization using non-individualized HRTF functions. They compared localization in the free field with localization using HRTF functions measured on a single good localizer. The listeners that participated in this experiment were inexperienced. One of the main findings of the study is the relatively high individual differences between the experiment participants. Of the 16 persons that participated in the study, 12 gave similar responses while the other 4 responded poorly with respect to stimuli locations. For the 12 participants who behaved similarly, localization in free field and

using non-individualized HRTFs is comparable with respect to azimuth judgements. With respect to elevation, judgements deviated more between the free field and the non-individualized HRTF conditions. The authors do not provide explicit information on the error rates however, from the figures error can be estimated to be at maximum 30°. For the participants with poor performance error rates were considerably higher and vary in a way that is difficult to quantify. Of considerable importance, is the particularly high rate of confusions that were observed in this study. Free field confusions were 6.5% in mean value but were increased by 3.8% for virtual sources for the participant group that performed well. For the participant group with poor performance confusions were 32.2% in mean with a maximum value of 43%. This means that participants were confused with respect to the direction of almost half of the stimuli. Similar confusion rates have been observed in pinnae occluded free field localization studies a fact that casts doubt on the validity of the mannequin's HRTF functions in this study. Front-back confusions were much less than up down ones. The results of this study indicate that non-individualized HRTFs maybe used for the creation of virtual spatial audio impressions, however the high rate of confusions and problems with modelling elevated sounds may hinder their applicability.

Wenzel and Foster [129] investigated the perceptual effects of interpolating non-individualized HRTF functions. They found the effects of interpolation to be relatively small compared to the effects of using non-individualized HRTF functions.

Wightman and Kistler [132] investigated whether head or sound source movements can help in alleviating the phenomena of confusions. Participants in their experiment were asked to judge the location of the sound stimuli, without moving their heads, moving their heads in a free style manner and moving their heads so that their nose was pointing to the direction of the sound stimuli. Individualized HRTFs were used and the conditions were replicated for both free field sounds played from loudspeakers and virtual sounds created by HRTF filtering. When listeners were asked to move their heads, the sound field was updated in real time based on information on the orientation of the participants' head. According to the results, head movements helped in reducing the frequency of confusions, however no major improvement in the localization accuracy was observed. The benefit was for both the real world and the virtual condition. It should be noted however, that the confusion rates were higher in the virtual case compared to the real. In a second experiment the authors showed that sound source movement that is not controlled by the listener does not help in alleviating confusions. This implies that in order to reduce the frequency of confusions, sound source movement must be controlled by the listener. The listener must be aware of the direction of the movement through other feedback mechanisms, such as proprioception in order to interpret the dynamic acoustic cues.

Due to the benefit of active listening, the process of using feedback from head movements to disambiguate confusions, further studies have been done to examine implementation parameters of this option. Wenzel [130], investigated the effect of system latency on localization accuracy in systems implementing active listening. In the experiment, system latency varied in levels of 33.8ms, 100.4ms,

250.4ms and 500.3ms. Latency refers to the time elapsed from the transduction of an event or action, such as the movement of the head, until the consequences of the action cause the equivalent change in the virtual sound source. The results indicated that localization judgements were accurate even when latency was at its highest value. However, there was a significant main effect of latency on localization error, with localization accuracy degrading as latency increased. Mean absolute error angle was 25° for a latency value of 33.8ms and ranged up to 32° for latency value of 500.3ms. Front-back confusions were minimal and unaffected by latency. Up-down confusions were affected by latency however planned comparisons showed that only the pairs of 33.8ms and 500.3ms differed significantly. Finally, latency had to be at least 250ms to be readily perceived by the participants of the experiment.

Brungart *et al.* [23] also investigated the interaction of head tracker latency, source duration and response time in the active localization of virtual sounds in the horizontal plane. They found that head tracker latency less than 73ms had no effect, neither in the speed nor in the accuracy of localization judgements. They found that for higher latency values, localization judgments with no time constraint were not affected. However, when participants had to respond within a given short time range, they consistently larger localization errors were observed. The authors also found that localization accuracy was improved by increasing stimulus duration and were thus able to recommend that designers should use long sounds in exocentric displays so that they can compensate to a certain extent for localization errors due to latency.

Begault [8] investigated the effects of synthetic reverberation on localization in virtual audio environments. He also used non-individualized HRTF functions. This study is particularly interesting because synthetic reverberation creates a richer user experience. Usually sounds in a virtual audio display are rendered as emitting in the free field; however, this does not correspond to real life experience. In this sense it is interesting to examine whether reverberation will create a positive effect on localization judgements. Reverberation was created by using two floor reflections, 64 early reflections from walls in the room and a late reverberation using noise. Reflections were HRTF filtered depending on their direction of arrival. This study used sounds that were located in the horizontal plane at eye level (0° elevation). For most of the participants, reverberation did not help alleviate front-back confusions, which remained at a rather high level of about 33%. Absolute azimuth judgements were significantly worse in the reverberant case compared to the free field case. Mean absolute deviation from target was 11.9° for the dry soundscape and 22.9° for the reverberant one. In addition, responses were more dispersed in the reverberant case with standard deviations being higher. A phenomenon that was observed in this study is that judgements of sound direction were biased towards the sides. In addition, most participants tended to perceive sounds as elevated at a mean elevation value of 12° for the dry and 23° for the reverberant case. Reverberation was found however, to help alleviate the problem of intracranial localization. Reverberant sounds were significantly less susceptible to intracranial localization. It is worth noting that intracranial localization was nearly eliminated in the reverberant case while it was quite evident for some subjects in

the free field case. In conclusion, when elevation is used in a virtual audio display that uses non-individualized HRTF functions, an improvement on the externalization of sounds is observed however this is to the detriment of localization accuracy. This can be explained by the effect of non-individualized HRTFs that is probably amplified when reflections are also filtered by them.

In another study by Begault *et al.* [10], the impact of head tracking, reverberation and individualized HRTF on sound localization was investigated in a single experiment. The stimuli in the study was speech, a more realistic stimuli compared to Gaussian noise or trains of Gaussian noise pulses that are commonly used in psychoacoustic experiments. Stimuli ware positioned at 0° elevation for six positions in azimuth distributed in a circular area around a user at eye level. The choice for 0° elevation was made because of the speech stimuli that were used. Apparently, speech stimuli do not contain sufficient energy in the high frequency spectrum that is thought to contribute to elevation judgements. The experiment compared anechoic presentation of stimuli to presentation using early reflections, and full auralization that contained early reflections and room reverberation processing, under presentation using individualized and non-individualized head related transfer functions, in a head tracked or non head tracked acoustic environment. An interactive graphic of a head was used by listeners to indicate the position of the speech stimuli in each trial. Participants oriented the graphic using a mouse so that it faced the perceived direction of the sound event. Although stimuli were constrained in elevation, the graphic could be oriented with respect to elevation. No effect of head tracking or HRTF type was observed for azimuth localization errors. A significant main effect of reverberation was, however, observed. In addition, when non-individualized HRTFs were used, head tracking was found to improve localization accuracy. Mean absolute localization error was approximately 23° for anechoic, and around 16° for the reverberant conditions. No effect of head tracking or HRTF type was observed for elevation localization errors. A significant main effect of reverberation was observed, which however was in the opposite direction compared to azimuth errors. Reverberation was found to increase the magnitude of the mean elevation error. Mean elevation error was 28.7° for the reverberant compared to 17.6° for the anechoic conditions. Head tracking was found to significantly reduce front back confusions, with mean values being 28% for the head tracked compared to 59% for the non-head tracked conditions. Participants in this experiment were asked to provide distance judgements that were used to judge the externalization of the sound event. There was a significant main effect of reverberation on the externalization of sounds. On average, participants felt sounds were coming from a distance greater than the one of their heads at a rate of 79% under reverberant conditions compared to a 40% under the anechoic conditions. No significant main effect of head tracking or any other treatment was observed with respect to externalization error. The fact that individualized HRTFs did not increase localization accuracy was attributed by the authors on the relative small variation of interaural time differences when comparing the ones measured and the ones represented by the model HRTF.

2.5 Discussion

The psychoacoustics literature review provides useful insight on the potential of 3D audio as a display modality. The literature verifies that virtual spatial audio systems can convey a spatial audio impression that is comparable to, however slightly worse than the real world directional hearing experience. The main problems with virtual audio systems are localization error, confusions of direction and externalization of sound. The two first problems also appear in the real world, however their effect is more pronounced in virtual systems. In the following text, a number of hypotheses of the effect of the problems in the perception of direction on interaction will be outlined together with potential solutions to overcome the problems. In the thesis, distance and motion are not considered explicitly as design tools, due to the fact that they cannot be judged consistently and that the thesis focuses on interaction with the direction of spatial audio elements. For this reason no speculation on their potential and their shortcomings is provided.

The discussion shows that localization error depends on sound direction. This in effect will result in different display elements being localized with higher ambiguity than others. The solution to this problem is to either compensate by design to provide a uniform target size or work with variable display element sizes. The problem of confusions is expected to influence performance in the display to a great extent. The problem becomes more pronounced because most of the commercial spatial audio systems use generic HRTFs. This is due to the fact that there is no widely available system for individual HRTF measurements and, in addition, because of the fact that the process of HRTF filtering is laborious and requires specialized equipment. With generic HRTFs confusion rates can be up to 40%, indicating that certain users will become confused on the direction of the display elements in nearly half of their selection attempts. Such a fact is unacceptable from a usability point of view. For this reason any system which uses non-individualized HRTFs and tries to use elevation and front-back cues without compensating for the confusion problem is expected to face serious usability problems. With individualized HRTFs the problem becomes less pronounced so it can be recommended that systems using elevation and front back cues should be relying on individualized HRTFs.

Sound source externalization is a requirement for spatial audio systems because although it does not affect perception of direction, it can result in annoyance and an unnatural feeling. To overcome the problem, special considerations with respect to the reproduction system have to be made. For headphone reproduction or reproduction using earphones, the system has to compensate for the frequency response of the transducer. Using open headphones, or headphones mounted at a distance from the ear are expected to further help in alleviating the problem. There are commercial solutions for portable reproduction systems that use loudspeakers. One solution is the shoulder mounted loudspeaker system used in [103]. However, social and privacy issues make such a solution less desirable than headphone reproduction for systems that are to be used in a mobile context. It should be noted here that headphone reproduction has

the problem of alienating the user from the natural audio environment. Such a problem may be overcome using bone-conductance headphones or acoustically open headphones.

In order to proceed with display design, the first decision that has to be made is regarding the selection of the display space. Based on the data on localization error, the area of the most accurate perception of direction is the area that lies in the front and on the horizontal plane. In this area, perception of direction is dominated by azimuth perception and is therefore based on interaural time and intensity cues. In addition, perception of direction in this area is invariant to problems related to non-individualized transfer functions due to the fact that interaural differences depend on head size, a factor that varies substantially less than outer ear shape. By informing the user on the display area it is also possible to overcome the problem of confusions. This is due to the fact that knowledge that the display area is restricted to the front will result in the user focusing on azimuth cues and ignoring cues related to elevation. For all these reasons, display design in the frontal horizontal dimension is the most conservative and safe choice, which is expected to minimize the negative influence of the uncertain nature of auditory localization.

The literature also provides insight in sound design. Based on the review, it is possible to recommend that designers use broadband sounds with substantial variation over their time course to enhance localization ability. However, for a display area that is constrained according to the previous paragraph, this constraint can be relaxed due to the fact that azimuth perception is more invariant to spectral cues compared to elevation perception. However, narrowband signals should be avoided to decrease the likelihood of elevation misperceptions and reduce the cognitive demand on the user.

In the aforementioned ways, many of the problems related to the perception of direction are alleviated. What is important, however, is to examine whether virtual spatial audio systems can support aimed movements. Although not directly referred in the literature, it is evident that audition does not provide a way to separate the boundary between the target and the background. This is to a certain extent due to localization error but in addition due to the fact that audition does not explicitly carry boundary information. In this sense, it can be argued that there is lack of support for corrective submovements that lead the user on to a target. Any effort to restrict the user onto a certain target area is due to face problems, especially when considering a scalable system that should support a large number of display elements. It might be necessary therefore to explicitly mediate target size information to the user.

2.6 Conclusions

In conclusion, the literature review indicates that there are a number of shortcomings in the perception of direction that have to be overcome to support interaction with a spatial audio display. Most of these shortcomings are also present when perceiving sound direction in the real world however their magnitude is less. These shortcomings can be summarized as: localization error, confusions and the use of non-individualized HRTFs. Localization error causes ambiguity when judging sound position, with

listeners not being able to perceive with accuracy the exact position of the sound. In addition, the sound signal cannot successfully convey information on the dimensions of the sounding object and the space it occupies.

Problems with modelling elevation and front back spectral cues will essentially result in higher values of localization error for elevated sounds and confusions of sound direction. The ratio where these confusions appear is critical. As was discussed in the chapter when using non-individualized HRTFs, confusions can be in the region of 40% for untrained listeners, a fact that greatly diminishes the utility of these cues. Using individualized HRTFs can help to overcome this problem however, such an option is not available for the mass population and in addition the problem is not eliminated.

With respect to Research Question 1, it is therefore evident that in order to support spatial audio target acquisition it is necessary to overcome the problem of localization error and the problem of confusions. Overcoming the problem of confusions can be achieved by constraining the display area that can be used effectively to the frontal horizontal plane where perception of azimuth is uniform across people. Localization error is expected to result in ambiguous feedback with respect to the exact sound location and the space it occupies in the display (target width). This limitation is critical because detailed information is necessary in order to support aimed movements as will be shown in Chapter 4 and it appears that feedback will have to be provided to compensate for it.

In the thesis, only the effect of the shortcomings on target acquisition is examined. As has been mentioned earlier, issues related to the presentation of sound sources in the display area are very important in the context of direct manipulation. However, the complexity of the issue is such that it can not be considered appropriately in the thesis. In the experiments presented in the thesis an effort is made to overcome target acquisition shortcomings by design, by constraining display space and by augmenting the selection task. For this reason a number of alternative selection techniques are evaluated experimentally with the goal of becoming able to design a usable selection task.

3 A Review of Existing Research in Spatial Audio Displays

3.1 Introduction

This chapter provides a summary of applications that have used spatial audio and an overview of the design choices that have been made in these applications. As was mentioned in Chapter 2, spatial audio systems essentially succeed in presenting a sound as appearing to emit from a certain direction in space. This is achieved by filtering the sound signal using filters that capture the properties of the transmission path between the location of the sound emitting object and our inner ear. These filters are different for each individual and the technology results in sound localization performance with similar accuracy as in the real world. Most current systems, however, use non-individualized filters and as a result localization accuracy is slightly degraded and in addition the problem of confusions appears. The term confusions refers to the case when listeners perceive a sound that is emitting from the back as emitting from the front and vice versa or a sound that is emitting from below as emitting from above and vice versa. Most of the spatial audio systems work using non individualized HRTF filters and the phenomenon of confusions can appear at high rates especially for non trained users. Application designs that use spatial audio make use of systems to present sounds in the display as emitting from a certain direction in space. The result is what has been called a spatial audio display. Spatial audio displays are coupled with a control mechanism to result in interactive systems where information is presented using the audio channel. When the application domain requires it, the location of the sounds in spatial audio displays is updated in real time with respect to listener position in the real or a virtual world, enabling what has been called the active localization of virtual sounds. Spatial audio systems may reproduce sound either through headphones or loudspeaker presentation.

Many of the studies in spatial audio displays do not build on any design rationale, they are rather ad hoc proposals and quite often no or only informal evaluation is made. Some of the studies however, do provide formal evaluation with respect to certain display aspects and their results can be generalized.

Based on the literature review, the chapter proceeds to identify particular design aspects common in the literature and brings them together to help the reader acquire an overview of the state of the art in the field focusing in:

- design choices with respect to display architecture
- reproduction options
- the control mechanisms that are used to enable interaction with spatial audio displays
- the application areas where spatial audio has been used

- implementation issues

A large number of the applications that will be presented are based on the concept of Audio Windows, an attempt to transfer direct manipulation principles into the spatial audio domain. However, the potential of the audio modality to support this kind of gesture interaction has not been assessed in the first place. For this reason, the chapter concludes with an investigation into the underlying cognitive and perceptual factors that support direct manipulation and identifies the requirements that the combination of control input using gestures and presentation through the audio modality should comply with, in order for auditory direct manipulation to work.

With respect to the Research Questions outlined in the Introduction, this Chapter contributes by identifying that they are novel and by providing insight on the direct manipulation requirements it hints into ways they might be answered.

3.2 Review of Application Designs based on spatial audio

Research in spatial audio displays is mostly concerned with providing designs and ideas about how people could interact with spatial audio sounds in certain application areas. Most of the designs are in accordance with the 'Audio Windows' concept that was proposed by Cohen and Ludwig [28]. The 'Audio Windows' concept refers to the application of the direct manipulation principles in the audio domain based on the ability of the human auditory system to perform directional listening. The feasibility of the concept is examined in accordance with the direct manipulation rationale in Section 3.8.

3.2.1 Handy Sound & MAW

Cohen [27] presented two applications based on the audio windows concept, Handy Sound and MAW. In Handy Sound, Cohen provides a framework for gesture interaction with spatial audio. According to the design users wear a DataGlove that has a Polhemus sensor on it. In this way data on the position and posture of a user's hand can be obtained. Using this information, interaction with spatial audio sounds is performed. This is done by physically manipulating sound source position. Users could grab a sound and move it around in space and then release it in its new position. In addition, users could point to a sound to bring to the foreground and perform finger gesturing to highlight it. Two finger postures were also used, one with fingers extended and fingers closed. The first highlighted a sound whereas the other cancelled the highlight effect. Depending on the current gesture, appropriate audio feedback was provided to the user. Three types of audio feedback were used: spotlight, muffle and highlight. Spotlighting was applied when the user was pointing to the sound. The aim of spotlighting and highlighting was to increase the brightness of the target sound however, no implementation details are given. Their difference was that highlighting is persistent, whereas spotlighting is only applied when a user points to the source. Muffling was implemented as a low pass filter.

Cohen suggested that feedback can be given by filtering operators that when applied to a target sound, change its perceived qualities. Cohen named these perceptual operators 'filtears'. Cohen considered filtears as tools to embolden or italicize a sound in the display. For example, making an audio element sound as a whisper could provide an indication of a certain auxiliary status of a sound if applied to the voice of a trusted advisor in a teleconferencing scenario. In addition, the 'filtears' are used to highlight or muffle a display element. However, Cohen does not provide proposals on the implementation of filtears, rather he suggests that they might be one out of the digital filtering possibilities that exist in filtering a sound without altering its timbre or degrading its intelligibility.

MAW is a more applied version of the 'Audio Windows' concept in the context of teleconferencing, In the MAW design, participants in a teleconferencing scenario are given a position in space and their voices are spatialised. A MAW user is assumed to be sitting on an orientation tracked chair that denotes his or her orientation relative to the speakers, and also to be wearing a DataGlove. In MAW the orientation of the speakers is updated in real time with respect to the orientation of the user, resulting in what is otherwise called an exocentric display. In addition, hand gesturing is enabled that allows the user to interact with the sounds in a manner similar to Handy Sound. A user can apply a certain sound effect to a sound and reposition the speakers to his/her own satisfaction. None of the systems proposed by Cohen were evaluated. Cohen's work forms the basis of spatial audio display design by introducing gesture interaction with 3D audio and feedback. His proposals are very interesting but a large number of issues related to the usability and the design of such systems are not addressed. In particular, the suitability of spatial audio in supporting direct manipulation based interaction is not examined. However, this early work is the basis of all research in spatial audio display from the time of its publication onwards.

3.2.2 Spatial Audio & Teleconferencing

Baldis [7] presented a study on the effects of spatial audio positioning of sounds, on memory, comprehension and preference during desktop tele-conferences. In this study, three different settings were examined: a non-spatial, where all conference members voices were played through a single loudspeaker in front of the listener, a co-located where four loudspeakers were used located at (-15°,-5°,5°,15°) azimuth and one scaled, where loudspeakers were placed at (-60°,-20°,20°,60°) azimuth. As can be inferred no virtual spatial audio system was used, sound location was communicated through speaker positioning. Speaker identification was better in the scaled, followed by the co-located and finally the non-spatial condition. Spatial presentation made it easier for participants to determine which conferee was speaking, requiring them to pay less attention to the conferee who was speaking and resulted in an overall better impression of the conference. In addition, spatial audio enhanced their comprehension of the conference and was preferred to non-spatial presentation. The benefit of using spatial audio was in general increased with increased spatial separation.

3.2.3 Head Mounted Spatial Audio Pie Menus

Of considerable interest is the study by Brewster *et al*. [19]. They investigated different design options for the creation of a spatial audio display for mobile/wearable computing that was controlled by the head of the users and the gesture of nodding. The idea was that users could nod to the location of sounds in the display to select them. Three competing designs were compared: an egocentric, an exocentric and an exocentric periodical. All displays featured four sound elements. In the egocentric case sounds were positioned at the four cardinal points with a separation of 90°. In both exocentric displays all sounds were positioned in front of the user at intervals of 40°. In the exocentric cases, there was support for a backwards nodding gesture that repositioned the sounds in front of the user when he or she turned for 180 degrees. Audio feedback was given when users were crossing boundaries between allocated active display areas. In the exocentric periodic interface, sounds were played one after the other, whereas for remaining two designs sounds were playing simultaneously. Error rates were in the order of 20%. Brewster *et al*. mentioned that their system could be used for navigation in hierarchical audio menu structures while mobile. In the evaluation users had to walk 20m laps around obstacles set up in a room in the university while interacting with two egocentric and one exocentric display, where selection was performed by head nodding. With respect to time taken to select, the egocentric display was significantly faster than the rest of the displays followed by exocentric and exocentric periodic. In relation to this, participants completed the tasks in significantly fewer laps in the egocentric compared to the rest of the displays. In addition, it was found that when using the egocentric display participants were able to walk at around 70% of their preferred walking speed, while maintaining half of their walking speed in the rest of the conditions.

A similar design for static users was proposed by Crispien *et al.* [30]. The proposed a direct manipulation type of display that featured spatial audio elements positioned at 30° intervals. A maximum of four elements would be played depending on the orientation of the user's head. Users would interact with the system using gestures and speech recognition. However, their system was at a development stage and no evaluation results were presented.

Finally, a pie menu proposal for hierarchical menu navigation was the one by Cooper and Petrie [29]. They proposed to arrange sounds according to a metaphor that resembles the solar system where users could locate elements of interest and 'converse' with them in order to achieve their goals. Their system was however under development at the time of its presentation and no evaluation results on the feasibility of the idea exist.

3.2.4 Grid Menu

Another proposal of an interface for audio hierarchical menu navigation is the one by Savidis *et al.* [102]. In this study a proposal for a hand gesture interface that supports pointing to 3D audio targets was presented. The system assumed a seated user who was pointing to 3D audio items that were located in a

grid front of him or her. Unfortunately no evaluation or design recommendations were provided by the authors.

A large number of designs for spatial audio interfaces come from the MIT group of Chris Schmandt. The work from this group is mainly addressing interaction with spoken information, such as news streams, documents, radio and phone calls and messaging. The designs are using spatial audio to assist simultaneous presentation of multiple audio information streams and gestures for interacting and navigating in the spatial audio environments.

3.2.5 AudioStreamer

The very first of these designs is Audiostreamer [105]. Audiostreamer is an application designed to enable listeners to attend to three spatialized simultaneous news streams positioned at -60°, 0° and 60° in azimuth, at the level of the user's nose. News streams were amplified or attenuated based on inferences of user interest towards each of them. Amplification was proportional to user interest but it decayed with time to enable listening to the rest of the news streams and to reflect a potential diminishing of user interest towards the news stream. User interest was calculated by the number of head movements towards an information stream. Head movements were tracked by a head tracking device. Three levels of interest were defined. At the highest level, only the news stream in focus was audible and the time before the amplification started to fade was maximum. At the other two levels a 20dB and 10dB gain was applied to the desired stream and the fading started faster. The news streams were tracked to reveal story boundaries. On such an event, users were notified by a sine tone while at the same time the level of the associated news stream was amplified by 10 dB. This application shows how the spatial audio's support for attending simultaneous information streams is used in a context different from teleconferencing, the one of simultaneous attending of news streams.

3.2.6 Dynamic Soundscape

Spatial audio has also been used for browsing textual information as in Dynamic Soundscape by Kobayashi and Schmandt [62]. Document contents were presented using synthetic speech, as if metaphorically spoken by a speaker. The length of the document was mapped to a circular orbit around the head of the user. In this way, apart from content, an indication of content's context was also provided by the location of the speaker relative to the user's head. Simultaneous speakers could be initiated by pointing at the desired location on the speaker orbit. In this way different parts of the document could be accessed simultaneously. Three control options were used: a touchpad, a knob and a gesture interface. Gestures or the control device could be used to initiate a speaker at a certain location on the orbit and also grab and reposition a speaker. To assist intelligibility, different timbres were used for each speaker. The system was not evaluated.

3.2.7 Audio Hallway

Spatial audio has also been used for the creation of virtual audio environments as in the Audio Hallway by Schmandt [104]. In this application, a virtual 3D audio hallway that was augmented with side rooms was implemented. The application was targeted to news browsing. The idea was to use the each of the rooms off the hallway to present detailed news excerpts of a common category, while a user moving along the hallway could obtain an overview of the content in each room, in a fashion similar to a user eavesdropping. Movement in the virtual world was controlled by the user's head. By moving their heads upwards or downwards users were able to move back and forth in the hallway, velocity being proportional to the head degree motion. The system was designed so that there was certain overlap between the rooms so that the audio overviews from adjacent rooms were concurrently audible to enhance the browsing experience. Audio overviews were a braided version of the whole story. The hallway was designed so that three audio overviews were heard at maximum. One was played in one ear and the rest on the other ear, one displaced to the front and one displaced to the back of the user. Users could enter a room of interest by nodding their head towards the direction of the room. When inside a room users were able to hear the content of up to eight stories, only four of which could be audible at the same time depending on the direction of the user's head. The stories were presented in a radial pie menu in front of the user's head. In order to increase the separation of the stories, a fish eye approach was used. As the user's head rotated a virtual lens moved across the audio sources so that a small movement of the head resulted in a greater distorted movement of the sources. Informal evaluation revealed that the hallway part of the system was difficult to use, the most important problem being that users could not browse effectively when moving at high speeds along the hallway. However, most users were comfortable with browsing stories once inside a room.

3.2.8 Nomadic Radio

Spatial audio has also been used in the context of a mobile application in Nomadic Radio [103] by Sawhney and Schmandt. Nomadic Radio is an application targeted at supporting messaging in mobile contexts. Output in Nomadic Radio is displayed using spatial audio reproduced by shoulder mounted loudspeakers and input is done by a command vocabulary through a speech recognition interface. Nomadic Radio offers two modes of interaction with messages, navigation and notification. Notification utilizes filtering and prioritization techniques to determine the timely nature of information and notify the user in an appropriate manner. Notification level is dynamically scaled and adapted by inferring interruptability based on user's recent activity and context of the environment. The system utilizes an algorithm to decide on the priority of an arriving message based on a variety of information mainly available in the user's calendar, contacts and to do list. This information is combined with usage level and an estimation of the likelihood the user is engaged in conversation to decide on an appropriate notification mode for incoming messages or phone calls. There were seven notification levels that constituted a

scalable auditory presentation method. Audio messages were formatted according to current notification level. The formats were silence (i.e. no presentation), ambient cues, non speech cues, synthesized speech previews, voice excerpts and foreground rendering based on a spatial mapping designed to catch user's attention. Navigation allows users to actively browse messages via a synchronized combination of non-speech audio, synthetic speech and spatial audio techniques. In navigation mode, the user was able to scan through his messages with the aid of a spatial layout. Each message was played at a different location in a circle around a user so that its time of arrival was denoted, with a certain overlap between inter-message onset times. The presentation level was set by the user out of the ones available by the supported scalable auditory presentation implementation. Informal evaluation showed the system to be usable, however the command vocabulary was in certain cases difficult to remember. For this reason, a preview of the available commands was also available. Nomadic Radio is relevant to the thesis since it uses spatial audio in a mobile context. Indeed, the authors used the eyes free interaction spatial audio enables in the design of an application for mobile users. However, control was not done using direct manipulation but rather through an agent based approach built on a command vocabulary using speech recognition.

3.2.9 Virtual Audio Guidance and Alert System for Commercial Aircraft Operations

Begault *et al.* [9] investigated the effectiveness of spatial audio for the guidance and alert of commercial aircraft operations. In an experiment participants were asked to acquire a visual target with th help of a 3D audio cue that sonified the position of the visual element and a monaural cue that included no information on the position of the visual target. They found that 3D audio improved the acquisition time of visual targets, however it did not affect the number of targets acquired.

Bronkhorst *et al.* [22] investigated whether a 3D audio display can succesfully convey directional information in a flight simulation scenario. They used individualized HRTF's in an exocentric design to sonify the direction and distance to a target aircraft. Participants were tested in a flight simulator that featured a display designed to assist search. The assistive display was either a visual display, a 3D audio display alone, the combination of the 3D audio and the visual display or no assisitng display. Participants were fastest when using both displays, followed by 3D audio, visual and no display. However, the differences among the 3D audio the visual and the combination of the two were not significant. The results indicate that 3D audio improves the search of visual tasks, however the results were not conclusive. A verification of the fact that spatial audio aids the acquisition of visual targets was provided by Bolia *et al.* [15].

3.2.103D Web Browser & WIRE

A more recent system using spatial audio was an interactive 3D audio browser by Goose and Moller [48]. In this system, spatial audio was used as a tool to support browsing of web pages. Presentation of audio information was done in the horizontal plane in front of a listener. Document

content was spoken out by a listener that was placed directly in front of the user. Three different voices were used for speaking out titles, normal text or links respectively. At periodic time intervals the current position as a percentage of the whole document length was announced by a female voice positioned appropriately upon a semi-circle in front of the user. When titles or links were encountered an earcon was played at an appropriate position along the semicircle in front of the user to indicate the title or link's position in the document. Although interaction was not explicitly considered in this design, a proposal for sonification of link traversals was provided. A non-speech sound moving from the original to the destination position in the document was used for this purpose. The sound followed a low trajectory in front of the user for intra-document links and was flying at a high trajectory for inter document traversals. The system provided the option for a fast-paced document preview or for selected presentation of formatted document content such as links, headings or structural grammatical information. The system presented in this paragraph formed the basis of a future system by one of the authors that used spatial audio to enable web browsing for people driving [46]. In this system a commercial radio interface was used, where stations corresponded to links to the user's favorite pages. The presentation was done by speakers positioned appropriately on the sides of the driver. For both systems no evaluation results were presented. Goose also studied how accuracy of localization varies along a semicircle spanning left to right in front of the user for egocentric presentation. As can be seen in Figure 9, after about 45° and 135° sounds tended to be perceived as originating from the sides of the user. No details on the experiment are given in the paper, however the results indicate that the effective area where is limited due to the localization problem.

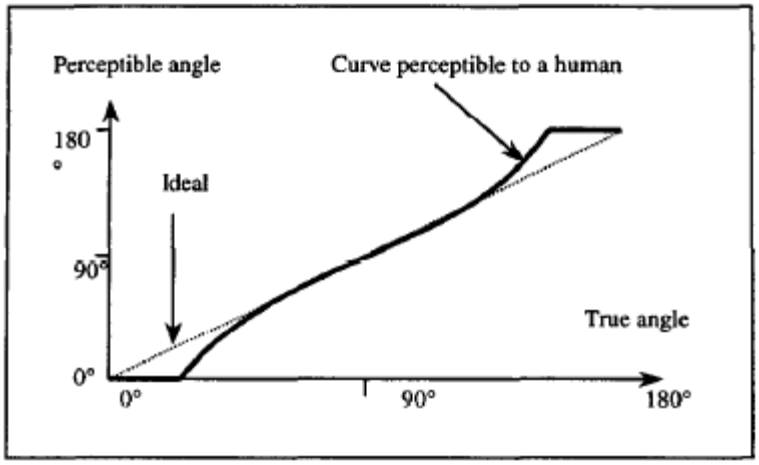


Figure 9. Localization accuracy as a function of angle in front of the user, 0 and 180 correspond to extreme left and extreme right. It is seen that people tend to perceive sounds as coming from the sides even when they originate from diagonal directions, adapted from [48].

3.2.11 Active Localization of Virtual Sounds

Loomis [68] performed a detailed evaluation of the active localization of virtual sounds. Active localization refers to the process users orient themselves towards a destination based on localization cues. The system was based on stereo cues and loudness update to communicate target position in space. Auditory cues were updated in real time depending on current user orientation and position. User position and orientation was tracked by cameras. Users were successful in getting close to the target, however they had problems realizing that they were close enough to the target. Distance update of sound sources was found to be enough to enable front-back disambiguation inherent in stereo panning. However, an elaborate model for distance modelling was used and in addition the field of navigation was constrained in a room. This study showed that audio cues are sufficient to guide a user to a target however no comparison was made to evaluate the effectiveness of navigation compared to other methodologies, such as sight.

3.2.12 Audio GPS

More recent auditory navigation systems rely on GPS tracking for user positioning. However, problems with GPS positioning, mainly due to noise, coverage and accuracy pose certain limits to the success of such systems. Holland *et al*. [57] developed AudioGPS. They used stereo panning to present target location and indicated front and back using two distinct sounds, a harpsichord for the front and a muted trombone sound for behind. The system could only discern user direction based on the position data gathered from the GPS device, because it did not employ a magnetometer to infer user direction. The sound design capabilities of the system enabled users to perceive eight positions in space, the ones at the four cardinal points plus four intermediate ones. However, when more precision was required, the user could match a ‘chase’ sound to the position of the navigation sound to achieve better accuracy. Distance to the sound was communicated by the repetition period of the navigation tone. When close to target, the tone repeated itself quickly but when farther away it repeated itself slowly. Alternative encodings of distance included repeating the navigation tone a fixed number of times numbers indicating an actual distance division. For example five clicks could be used to indicate 50m of distance. Informal evaluation showed that the system could be usable however; limitations in technology used hindered its usability.

3.2.13 GPS Tunes

Another attempt in the direction of audio assisted navigation was the one by Strachan *et al*. [116]. In this work an audio navigation system based on stereo panning was developed, however the audio stimulus was music and distance was communicated using the loudness of the music signal. This system is the first to provide audio guided GPS positioning in a handheld device. Informal evaluation showed that users were quite successful in using the system in the sense that they could reach the desired target.

3.2.14 Diary in the Sky

In Diary in the Sky [121] by Walker *et al.*, spatial audio was used to test the effectiveness of a potential mobile calendar application. A comparison was made with respect to the items recalled correctly after users browsed for the same time in the visual calendar interface offered in PalmOS and a spatial audio calendar display that was designed by the authors. In the spatial audio interface the events of a certain day were presented using speech and spatialized on a circle round the head of the user at the level of his or her nose at a position that was varied according to the time the event was programmed to happen. Participants responded to questions that were designed to test the recall of the day's events as a function of the presentation modality. Users performed significantly better when results were presented using audio, compared to when they were presented using the visual interface. In addition, the audio interface demanded significantly less subjective workload compared to the visual interface.

3.2.15 Monitoring Background Activities (Spatial Audio Progress Bar)

In a study by Walker and Brewster [122], spatial audio was used for monitoring background tasks in the (attentional) background. A spatial audio progress bar design was given as an example. The design consists of two sounds, a 'lub' component positioned in front of the user that is used as a reference, and a 'dub' component that is positioned on a circle around the user to indicate task progress. The transfer rate is encoded in the time difference between the presentation of the two sounds. When the task ends successfully two identical sounds are played in front of the user. In the case of a download stall only the 'lub' sound is played to indicate progress stall. An experiment was designed to evaluate the spatial audio progress bar design versus a standard visual progress bar design for participants immersed in a foreground visual task. According to the results, participants were able to monitor background task progress more accurately using audio, however response times were not significantly affected by modality. Users were able to progress faster in the foreground task in the audio condition compared to visual one. Overall, the spatial audio progress bar also required less subjective workload. In addition, this study showed that participants were more sensitive in the audio condition compared to the visual one.

3.2.16 AudioDoom

Of considerable interest is the application of spatial audio in audio games. Audio games are a relatively new class of computer games that use audio to display the game environment and enable interaction. In [70], Lubreras and Sanchez presented AudioDoom, a 3D first person audio game for blind children. AudioDoom works by splitting the navigable space into small atomic environments. Atomic environments define the minimal scenario of action in a given moment. While in an atomic environment, the child can interact with entities at different voxels, this is areas around it. The linear connection of the atomic environments was used to define a corridor. Navigable space is organized into several corridors, giving a semantic and argumentative connection of a hyperstory and the space. The corridors are

modelled as contexts and the doors as links. The child could perform a number of activities, to move forward to the next atomic environment by taking a step, to open a door, to make a turn and to interact with an entity in a certain way. Interaction with entities was performed by either picking up objects or shooting at enemy creatures represented with sounds. Evaluation of the game was performed by asking blind children to navigate and interact with the 3D audio environment and consequently build a representation of the environment using LEGO blocks. The results showed that blind children could sufficiently reconstruct the 3D audio virtual world they were experiencing in the game, indicating that spatial audio is an effective means of communicating spatial information also for the visually impaired.

As a summary, the design studies in the literature reveal that spatial audio systems are useful in tele-conferencing systems because they enhance intelligibility of simultaneous presentation. They also show that spatial audio can be successfully used to convey background information for users engaged in visual tasks. Interaction in exocentric interfaces was found to be slower that in egocentric ones. In addition, it is known that users have a problem homing to exocentric spatial audio targets and they cannot experience fast closure when asked to identify that they have reached an exocentric audio target. Finally, the perceptual impression of the direction of a sound that appears on diagonal directions is biased towards the sides of the user.

3.3 Summary of the Design Choices in Spatial Audio Displays

In this section, design choices in spatial audio displays are identified based on the literature review. The design parameter space can be segmented into the parameters related to the spatial arrangement of the display and parameters related to the presentation of the display. With respect to the spatial arrangement of the display the relevant design parameters are each element's orientation, distance and motion. All these parameters can be defined either relative to the user (egocentric design) or relative to the world (exocentric design). These aspects can be seen individually for each display element as well as in relation with the rest of the display elements.

Other design parameters include the semiotic mappings that will be used for display presentation, which can be in the form of auditory icons [43], earcons [13] and speech [99]. Although this aspect of display design is not within the thesis context, a short overview of signs for auditory display is provided in Section 3.8.2. Finally, intra-display element differences can result as design choices, a fact that offers another field of choice for the designer. Such differences can be the amount and the nature of the spectral overlap between display elements, relationships between their inter-onset intervals and relationships between their individual properties such as their rhythm, pitch and register and the rest of the musical listening related properties. These should be considered from the point of view of display intelligibility in context with auditory scene analysis [16] but also in conjunction with aesthetic aspects. Due to the fact that the thesis is dealing with the design of pointing interactions and how these might affect the spatial

arrangement of the display, it does not explicitly consider design options related to musical or everyday listening and speech processing. Such options are important in their own scope, in the context of auditory scene analysis [16], and need to be individually considered. However, they are out of the scope of the thesis and apart from referring to them when relevant the thesis does not attempt to disambiguate the related issues.

3.3.1 The Effect of choices on the Spatial Parameters of the Display Elements

With respect to orientation, distance and sound motion, the design choices have to be made in a way that no intelligibility deficiencies are encountered in the display and interaction with the display elements is usable. The relative orientation of the display elements to the user can lead to serious problems if the phenomenon of confusions is encountered. As was discussed in Chapter 2, a major shortcoming in state of the art virtual audio systems is confusions with respect to elevation and front-back direction of a sound event. Attempting therefore to utilise such a cue will result in a large number of cases where users will be confused with respect to the actual direction of the sound, a fact that will result in increased interaction time and user frustration, especially when they are not familiar with the display. For this reason, elevation and front-back cues have been little used in applications based on spatial audio displays. Most designs are confined in the frontal or the whole of the horizontal plane. Although front/back confusions might appear in that case as well, informing users that elements lay in the frontal direction, forces them to utilize azimuth cues and pay less attention to the front-back cues.

The alignment of the sounds in the display is done in two main ways. One follows a circular approach where sounds are placed on the circumference of a circle at the level of the user's ears. In this design elevation cues are not used, but in some cases front back cues have been used as in the work of Brewster *et al*. [19]. It is the case however, that most of the studies are confined to the frontal horizontal plane. An optional design has also used the frontal horizontal plane however, sounds were placed in a straight line in front of the user [48]. The major problem with this design is that the distance of the sound to the user becomes larger as sounds appear on the sides and for this reason sounds to the sides are less loud than the ones appearing in front. Compensating for this effect will essentially result in the circular impression that has been discussed earlier in this paragraph.

The absolute orientation of the display elements also affects their spatial separation. The spatial separation of the display elements is a design parameter that is influential both in terms of interaction performance as well as in terms of display intelligibility. Literature on the Cocktail Party Effect [4] reveals that the intelligibility of audio display elements is increased with their spatial separation. In addition, the spatial separation between display elements defines implicitly the target width that could be allocated to elements. It can be inferred from these facts that when display elements are placed close to each other, intelligibility as well as interaction problems could emerge. It is therefore of interest to

investigate this issue in detail to understand the effects of the aforementioned parameters. Such an investigation is undertaken in the context of the thesis in Chapter 7.

Distance and sound motion are design parameters that have not been used extensively in spatial audio display design. With respect to distance, the reason for this is that it is hard to make absolute judgements of it, especially for unfamiliar sources, as was discussed in Chapter 2. However, its usefulness in implying the relative proximity to a sound event has and could be used in navigational studies, such as [116]. The situation with sound motion is similar. Increased localization errors have been observed in sound motion studies, a fact that implies that its usefulness might emerge from its attention grabbing potential rather than from its localization properties.

3.3.2 Egocentric vs. Exocentric Designs

Due to their spatial nature, orientation, distance and sound motion can be defined in two frames of reference, relative to world and relative to the user. If users and sources are considered mobile, the used frame of reference is also implicitly affecting how the display will appear in the other frame of reference. For example, if display elements are defined in a fixed position in the earth coordinate system, orientation and distance of a sound relative to a mobile user is in such a case changes as the user moves. On the other hand, if the sounds were defined fixed in the user's coordinate system, the result would be that they would appear as moving together with the user in the world coordinate system. These ends appear to define a design continuum that spans from exocentric to egocentric. In between, different design manipulations can be performed, for example a designer of an exocentric display might choose to display in real time the orientation of display element relative to the user, but not the distance as in [57] or *vice versa*. Such decisions are dictated both by the application area but also by the perceptual aspects relevant to the intelligibility requirements of the display.

Egocentric displays are more suitable for interaction that has a repeatable pattern, like interaction with lists or menus because display elements are always in a fixed position relative to the user and thus they are easy to remember and acquisition time is minimized. Therefore, they are quite suitable for displays designed to follow a direct manipulation type of interaction. Such a design is also particularly useful for mobile users. When mobile, users change orientation often and thus keeping the display elements fixed to them would help them to remember their positions and preserve their interaction patterns.

In exocentric displays, display element positions are updated in real-time relative to the user orientation and appear to be fixed to the world. Exocentric presentations are quite useful for assisting navigation tasks either in virtual or real worlds. The applications presented in Sections 3.2.7, 3.2.12 and 3.2.13 are examples of exocentric displays. Their main benefit is that the presentation in such displays implies that the audio information presented lies in the world outside the user and for this reason the

sound it emits comes from a different direction depending on the orientation and distance relation between the user and the sound emitting event.

In an exocentric display, the audio scene is updated in real-time based on the orientation of the user relative to the direction of the display elements. Such a technique is usually implemented using a head-tracking device that provides the orientation of the user's head and this information is delivered to the spatial audio engine, which updates sound positions in real time. As a consequence, a sound that is defined to be to the left of the user will appear as in front when the user's head is facing left. From a sound localization point of view, active listening helps to alleviate up/down and front/back confusions in sound direction. However, its effectiveness on localization accuracy is small if any, (see Chapter 2). It is the case however, that such a cue can theoretically improve accuracy if used in the appropriate way. For example, a user interacting in such a display can move until the sound appears to be in front. This action is quite beneficial from an accuracy point of view since sound localization is most accurate for sounds in front of a user. Improvements in selection accuracy could therefore be expected when such a technique is used. On the other hand, the need for continuously attending to the available orientation cues might result in selection time deficiencies. The strength of this effect can vary with the extent to which the user is required to use the real time updated orientation cues.

In an egocentric display, sound position is fixed to the user and thus in such a display one can expect that localization accuracy will vary with the position of the display elements. In addition, such a display is prone to confusions since no mechanism of overcoming this phenomenon is provided. For this reason, localization accuracy in such a display would be lower. On the other hand, due to the relatively simple nature of the display, interaction is expected to be fast, if it is not influenced by errors caused by the localization problems.

Finally, an egocentric display is 'light' in terms of computational requirements. In fact, there is not even the necessity of a spatial audio engine. Sound files, can be HRTF filtered off-line and then replayed as simple wav files. On the contrary, an exocentric display is computationally heavy as the sound scene has to be rendered at a very fast rate if the result is to be convincing. In addition, there is the requirement for user tracking so that relative position and orientation towards the display elements is monitored. In a virtual world such an issue is not necessarily a problem. However, in the case of applications for mobile users, there are limitations in commercially available tracking devices (such as GPS) that might limit user satisfaction. However, when considering the implementation of such schemes in low computational power mobile devices limitations emerge. One solution to such a problem has been proposed by Goose et al. [47, 92]. The idea is to use the expanding network infrastructure to stream audio to the devices which will then act as content reproducers. Such an option is an interesting one to consider when the update rate required is not limited by the bandwidth and performance of the network infrastructure. However, it has to be evaluated in order to understand the limits of such an implementation strategy.

3.4 Reproduction

Decisions on the reproduction options in spatial audio displays are largely made based on the target application area. Reproduction using loudspeakers is a viable option for users that are expected to interact with the display outside a formal social context and where user mobility is confined to small areas. Possible application areas would include home entertainment systems, driving situations and office environments where only one person would interact with the system. Alternatives include loudspeaker systems that can be mounted on the shoulders of the user as is discussed in Section 3.2.8. Although such systems allow the user to move freely, the problem of operating them in a social context has not been solved.

Headphone reproduction is a viable alternative and its applicability is high due to the large number of people who use headphones, especially when on the move but also on other situations. Headphone reproduction provides good spatialization quality and the large number of headphone types available suit all user styles. However, the major disadvantage of headphone reproduction is the fact that it may occlude the natural audio environment from the user, a fact that is important when people move in the real world. The problem can be partly alleviated by using open headphones. A number of more promising alternatives include bone conductance reproduction and monaural presentation. A more thorough discussion on the issue is provided in the context of Experiment 3 in Section 7.3.

3.5 Control

The main control options that are found in the literature of interacting with elements in spatial audio displays are gesture control (either through physical movements or virtual pointers controlled using hardware devices, like knobs or mice), and control through a command dictionary recognized through speech recognition software.

3.5.1 Gesture Control

Gesture control is particularly relevant to the thesis since it forms the basis of direct manipulation interactive systems. Gesture control can be accomplished either solely based on pointing and dragging actions or based on a richer gesture vocabulary that can be used to provide higher level control options. The former case is the simpler and has been proven to be quite effective in direct manipulation systems, in particular in controlling the desktop metaphor. Gesture vocabularies could provide faster interaction. For example, to delete a display element a delete gesture could be performed while the item is in focus, instead of dragging the item in the recycle bin. However, remembering gestures from a vocabulary is a cognitive operation that requires frequent use and learning. In addition, gesture recognition systems require training to achieve high recognition rates for complex gestures, due to the fact that there is substantial variation in the way gestures are performed by people.

In this sense, one can speculate that as the size and the complexity of the vocabulary increases usability problems might emerge. On the other hand, simple pointing and dragging gestures are easy to recognise and are performed consistently for large classes of people. For this reason, the thesis focuses on simple gestures and examines pointing-based interactions. Another reason for this choice is that the elemental pointing-based interaction is the foundation of any direct manipulation system and needs to be resolved before more advanced gestures can be introduced into the system.

Finally, another argument for the using gestures in the thesis is that they can be easily performed when mobile based on kinaesthetic feedback, as is discussed in Section 8.3.

3.5.2 Control through a Speech Recognized Command Vocabulary

A proposal for spatial audio display control, that was done in the context of the Nomadic Radio application by Shawney and Schmandt [103], is that of control through a command vocabulary that is recognized by a speech recognition engine. The technique has its merits, however its main disadvantages are the fact that training is required for the speech recognition engine to work and in addition, the recognition rates might be not satisfactory when the system is operated in a noisy environment such as, for example, when a user is mobile. In addition, users might feel uncomfortable with such an option in a social environment. Finally, such a solution is demanding from a cognitive point of view since a user has to remember the control vocabulary at all times. The authors recognized this shortcoming and provided an option for command vocabulary playback in their system. However, such an option is not efficient in cases where the system is not used frequently since the lack of support for active exploration might hinder their usability.

3.5.3 The Control Option chosen in the Thesis

In the thesis context, gesture control is the most appropriate. This is due to the scope and the application areas in which the system is expected to function. The direct manipulation paradigm relies on fast and accurate pointing actions that are a form of gesture control. In addition, the application area the system is targeted to is mobile human computer interaction, due to the ‘eyes free’ nature of the display. Gesture control is a natural solution in mobile contexts due to the fact that gestures can be performed eyes free based on kinaesthetic feedback. The thesis opts for physical gestures rather than the stylus based control that has been commonly used in mobile contexts. The reason for this is that stylus based control has been found by Pirhonen *et al.* [93] to be harder to perform when mobile compared to control using physical gestures. In the literature pointing to audio items has been performed using hand (see Section 3.2.1,3.2.6,3.2.4), head (see Section 3.2.7,3.2.3) or virtual pointers (see Section 3.2.6). However, the different control options have not been evaluated against each other. In addition, the usability of spatial audio target acquisition using gestures has not been examined.

Gesture recognition is done through tracking devices. Such devices can track attributes such as position and orientation, as well as velocity and acceleration and this information can be used for recognising movement patterns that correspond to gestures. Such gestures can be as simple as the direction of pointing, but can also be comprised by more complex movement patterns. Until recently the technology behind such tracking devices required wired contact with base stations, a fact that greatly hindered their use in contexts such as mobile computing. Due to the fact that such technologies are used in the thesis a review is provided at this point that can be used for future reference.

3.5.4 Tracking Technology Alternatives

A tracking system that has been widely used is the Polhemus Fastrack [94]. It works using electromagnetic tracking and thus it functions even when obstacles exist between the device and the user. It provides six degrees of freedom, cartesian and rotational coordinates. However, it does require a base station and on top of that a cable connection between the device and the base station. Users can move however within reasonable distance from the base station, due to the large cable length; however this characteristic of the device makes it inappropriate for large scale experiments with devices on the move. On the other hand, the device is very accurate and reasonably stable. Another advantage is the fact that the device offers the possibility to control up to four trackers simultaneously, thus giving the opportunity to track several parts of the body at the same time. The Polhemus Fastrack system is depicted in Figure 10.

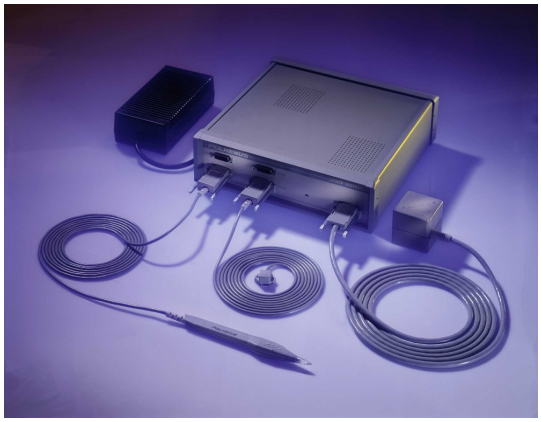


Figure 10. The Polhemus FastTrack System

Other tracking options include the Intersense Intertrax device [58]. This is a very small portable head tracking device, working by means of inertial sensing, that is able to provide 3 degrees of freedom that can map to the orientation of the head. A great advantage of the device is that it is USB powered and does not involve the use of a separate station as a reference unit. This means that the device could be used on the move, when appropriately plugged to a wearable computer. Unfortunately, the readings obtained

from the device are not stable to the point where its reliability is compromised. A picture of this device is illustrated in Figure 11.

Figure 11. The Intersense Intertrax tracker

Another tracking possibility is a device that can be worn on the user's hand such a virtual glove [101]. These provide information of the position and orientation of the user's hand as well as the posture of the fingers. In this sense they can be used in recognising grabbing gestures, pointing gestures and so forth. However, these devices also require a wired connection to a base station and for this reason although useful they are not considered appropriate in the context of the thesis. An example of such a device is the P5 glove that is displayed in Figure 12.

Figure 12. The P5 glove

Finally, a tracking device that was found to be accurate and reliable enough for the investigations that are undertaken in the thesis is the MT-9B tracker from XSENS [133]. The MT-9B is presented in Figure 13. It provides static accuracy of less than 1° and its dimensions are 3x3x3 cm.

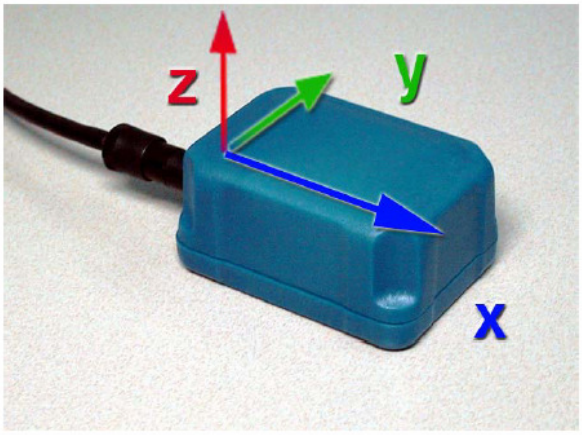


Figure 13. The MT-9B tracker.

This device works using a combination of inertial, magnetic and acceleration sensing and provides three degrees of freedom orientation measurements as well as acceleration and velocity estimations at a rate of up to 100 Hz. It can be connected to a wearable computer through a cable, is powered from batteries, it is very light and small enough to fit inside a user's palm, and does not require any base station. In this sense such a device is perfect for the experimentation that is done in the thesis and for this reason it is used in all experiments chapters apart from the first that is presented in Chapter 5. In the first experiment a Polhemus Fastrack device has been used due to the fact that the MT-9B was not available at this point. The Polhemus device has the reliability required, and due to the fact that mobility was constrained in the first experiment it provided a viable solution. For the rest of the experimental studies, in particular the one in Chapter 8, the mobility allowances of the Polhemus system would not suffice for the experimental task.

3.6 Application Areas

As can be seen in the review, spatial audio has been primarily used in three contexts: desktop computing, virtual reality and mobile human computer interaction. It was used either as the sole means of display output or supplementary to visual information display. In desktop computing, spatial audio has been used to enable users to effectively monitor background activities while occupied with a visual foreground task. The rationale for using spatial audio in this domain is the omni-presence of our hearing sense that alleviates the problem of occlusion that is common in vision. In addition, certain proposals have been made for the creation of auditory direct manipulation interfaces where interaction with spatial audio display elements is done in a fashion similar to the way it is performed in visual displays. Such systems enable users to interact with menu structures for example and also create audio desktops that can be manipulated based on pointing actions.

In the context of virtual reality, spatial audio has been mainly used to either enhance the experience of interacting with a visual virtual reality display but as well to create virtual auditory worlds. Virtual auditory worlds can be used as a means to present information in space but also to provide an environment for supporting a gaming experience. Spatial audio displays have also been proposed for mobile human computer interaction, because they provide a portable and eyes free way to interact. In this way, users can navigate freely in the environment while interacting with a system that where output is done using audio. The main application areas of spatial audio displays are presented in Table 1. It is worth noting that some of the applications were also targeting the visually impaired community, however this application area is out of the context of the thesis.

An issue of particular interest in all applications is sound design for display presentation. Application designers can choose between speech and non-speech sound, the choices not being mutually exclusive. This distinction is rather fundamental in audio display design and decisions on which type to support for each type of display output are primarily made based on the application domain and the interaction type that the display is designed to support. This issue is treated briefly in Section 3.8, since it is outside the scope of the thesis.

	Desktop	Virtual Reality	Mobile HCI
Teleconferencing	MAW - [27]		
Monitoring	[122]		
Direct Manipulation	Dynamic Soundscape [62] 3D Web Browser [48] Audio Streamer [105]	Audio Hallway [104]	WIRE [46]
	Grid Menus [102]		Pie Menus [19]
Audio Games		AudioDoom [70]	
Messaging			Nomadic Radio [103]
Navigation			AudioGPS [57] GPSTunes [116]

Table 1. A summary of application areas for Spatial Audio Displays

The number of application areas supported by spatial audio displays is limited. This is due to the fact that spatial audio reproduction technology has only been made available commercially in the past decade and became available for desktop and mobile computers more recently. Application areas considered so far include mainly, news, radio and document browsing, messaging, supporting navigation in real and virtual worlds and interaction with audio menus.

The designs are, with few exceptions, simple and cannot be thought to account properly for the design space. Spatial audio display design has not been examined in detail. No coherent proposals exist

for frameworks that can support application development. Most designers resort to a trial and error approach that potentially is not rewarding in terms of usability. There is a lack of comparative and evaluative studies to provide recommendations on design and indicate the benefits and shortcomings of interaction with sound. In addition, there is as yet no body of research on the appropriate areas for applying spatial audio in application design.

3.7 Implementation Issues

The implementation of a spatial audio display and its control is dependent on the application area the display is targeted to. In any case the main components of a spatial audio display are the audio spatialization engine, the control algorithms, the soundscape update system and the event handling mechanism. These subsystems can be computationally demanding due to the large number of real time operations that are necessary. However, it is possible to constrain the computational demand by design.

The least computationally demanding situation is that where the display is based on the playback of wav files which are presented in an egocentric manner. In that case, signal processing can be performed off-line and there is minimal demand on soundscape generation. The system however, should support a mechanism for changes in the position of the display elements, based on user demand, in which case re-filtering is necessary to present the display element from its new location. Such a system can be implemented in contemporary, low computational power handheld devices such the iPaq [56] or wearable computers.

In the case of a large scale exocentric system, however, where sound positions have to be adjusted in real time with minimal latency, faster systems such as laptops or desktop computers are required or dedicated hardware connected to a low computational power device. Unfortunately, such connectivity options are not readily available in handheld devices and therefore such an option is not common.

If, in addition, the soundscape is generated in real time using sound synthesis algorithms, then the computational demand is further increased. The use of sound synthesis algorithms in display generation can prove particularly useful since they provide options to adjust certain parameters of display elements in real time. For example, it is possible to alter the pitch or timbre of a sound or add reverberation to a certain element. Such options can be particularly useful for providing feedback as well as for minimizing the memory requirements of the display. Sound synthesis algorithms model sounds based on sets of parameters that require significantly less memory compared to storing a sound in a wav file format. In addition, the parameters are easily manipulated to change the timbre and pitch of a sound as well as its loudness. It is also quite common that the parameter space can successfully represent a large variety of timbres. Opting for a sound model for soundscape generation minimizes therefore memory demands and provides an easy and consistent solution for providing feedback. The downside is increased computational demands that might result in this solution being unfeasible in certain contexts.

To conclude, if the display is to be used in non-mobile situations, such as an office environment, then there is no particular limitation in computational demand. The extent to which spatial audio applications are superior to visual ones is however debatable and until research is published on cross modal comparisons no definite answer can be given. However, in mobile situations rich exocentric displays and sound synthesis algorithms are out of the possibilities of contemporary handheld systems. An alternative option in these cases is the one of wearable computers. Alternatively, The designer in such cases might wish to compromise display complexity to achieve a solution that works well in low computational power device. It is however the case that as technology improves more complex display arrangements will be possible to implement in handheld and wearable devices. This argument justifies research investigations into more complex spatial audio displays.

3.8 Identifying the requirements for supporting direct manipulation

Most of the applications that use spatial audio are based on the Audio Windows metaphor [27, 28] in which sense they are trying to apply direct manipulation in the audio domain. However, there is no scientific background on how and whether the direct manipulation principles could be applied in the audio domain. For this reason this section investigates the requirements for supporting direct manipulation and examines to what extent these have been verified to be supported by our hearing sense.

Direct manipulation emerged as a solution to the problem of how to present the interface affordances, mappings and constraints so that a valid conceptual / cognitive model is supported and interaction flow is not hindered. Interfaces to systems that are designed based on the manipulating and browsing interaction model are classified as direct manipulation interfaces [111]. Direct manipulation is a term coined by Shneiderman in 1982 [113] to describe existing successful systems. According to Shneiderman, '*these systems all had rapid, incremental and reversible actions, selection by pointing and immediate feedback (100-millisecond updates for all actions)*'. An integrated portrait of direct manipulation is given in [111] and it includes:

- Continuous presentation of the of the objects and actions of interest with meaningful visual metaphors
- Physical actions or presses of labelled buttons instead of complex syntax
- Rapid incremental reversible operations whose effect on the objects of interest is immediately visible

The objects and actions of interest in direct manipulation interfaces are rarely the actual system entities. Instead the interface builds on metaphors used to create virtual objects on virtual platforms in a convincing manner. The metaphor is the basis of the conceptual model the user forms of the system as this emerges through the perception and interpretation of the perceivable aspects of the interface. In other words, the conceptual model is of symbolic nature. Interaction is performed in the symbolic world that

according to direct manipulation directives must be designed according to familiar paradigms so that it supports familiar actions.

According to direct manipulation directives, users interact with the virtual objects in a way akin to the one used to manipulate similar objects in the real world. The most fundamental interaction modes are therefore analogies to physical actions used in the real world such as pointing and dragging and dropping. Using these interaction modes, it is possible to denote user interest on a certain interface element as well as to rearrange the metaphor space to suit the individual needs. Of critical importance to the success of the design is the provision of feedback. The role of feedback is to close the loop between the metaphor space and the users and to allow them to perceive the state of the display as this is emerging out of the outcomes of their actions. In this way it enables users to accomplish their goals.

The success of direct manipulation is to a big extent due to that it enables people to deal with task execution problems at the perceptual, rather than the cognitive level, by taking advantage of perception action mappings indicated by the environment rather than asking them to perform analytical problem solving [100, 119]. This is due to the fact that perceptual processing is fast, effortless and proceeds in parallel whereas analytical problem solving is slow, laborious and proceeds in a serial fashion. Such observations came out as the results of experiments where participants dealt with the same simple problems in either a perceptual or a cognitive way. Answers that resulted from perceptual processing were close but not exactly correct with performance exhibiting small standard deviation [106]. On the contrary, when analytical thinking was used, exact answers appeared more often, performance however (i.e. both in terms of time to arrive at the answer) exhibited much higher (ten times more) deviation. This is because analytical thinking can lead to big mistakes if a wrong path to solution is taken, however perceptual criteria depend on observations and perceived analogies that are experienced in a consistent way across groups of people as psychophysical research indicates. Consequently, behaviour based on perceptual and cue-action mapping skills can be to a certain extent predicted for a wide class of the population whereas the same is not necessarily true when analytical problem solving is concerned. Furthermore, behaviour at the perceptual level becomes automated by repetition thus resulting in a highly efficient way to deal with problems that reappear.

This rationale is behind modern interface design, in the sense that it encourages exploratory learning and goal accomplishment based on perceptual cues that guide action. Exploratory learning refers to the process of learning a system's functionality by exploring its functions. The conceptual model of the interface emerges as the result of a successful exploratory learning process.

Based on the aforementioned discussion it is possible to pinpoint the most important human abilities that are used in interacting with direct manipulation interfaces:

- space perception that enables us to differentiate between display elements
- motor skills that enable us to integrate external as well as kinaesthetic feedback to reach a desired position in space
- the ability to use semiotic information to create conceptual models assisting us to comprehend an interactive system and form appropriate goals
- the ability to engage in a closed loop and interpret the feedback provided by the interface to evaluate our actions and validate our conceptual model of the system

Space perception enables us to perceive the different spatial positions that have been allocated to the different display elements and associate the perceived display elements in relation with their spatial location. Motor skills, complemented by external (e.g. visual) and (internal) kinaesthetic feedback enable us to reach a desired location so that we can acquire and manipulate an element of interest. Our abilities in comprehending and producing signs enable us to infer the association of virtual objects in the metaphor space to their functionality and subsequently decide on the relevance of the objects with respect to our current goal. The ability to act in a closed loop mode of behaviour results in an engagement with the system and facilitates an exploratory mode of learning.

Most direct manipulation interfaces are visual. This is due to the fact that our visual sense is self sufficient in terms of space perception, adequate in supporting aimed movements and quite powerful in depicting and capturing symbolic representations of objects. However if one considers the aforementioned requirements of direct manipulation interfaces it is possible to infer that these skills might also supported by audition [85]. This fact led to the development of auditory direct manipulation which was conceptualized under the notion of 'Audio Windows'. The next section presents the concept.

3.8.1 Auditory Direct Manipulation

Based on the directional nature of hearing and our ability to process auditory symbolic information Cohen and Ludwig proposed Audio Windows [69]. Instead of spatially positioned icons Cohen & Ludwig proposed to use spatially positioned sounds. They proposed manipulation of the audio elements to be done using pointing based manipulations either through physical gestures or by using virtual audio pointers controlled by a hardware device. The finally proposed that feedback should be given using perceptual audio operators that slightly change sounds as a result of a certain signal transformation. According to Cohen *'the idea is to create a just noticeable difference', an acoustic enhancement that is noticeable but ignorable, unambiguous but unintrusive'* [27]. According to the proposal such feedback can be provided by filtering, echoes, reverberation or equalization.

The proposal of audio windows was associated with a teleconferencing scenario. The application of the concept in a more general context that will include the use of iconic information has not been evaluated. This is a major issue considering that direct application of the principles of visual display

construction in the audio domain is often not feasible and results in unusable systems [63]. The reasons for this can be sought in the differences between vision and audition. These differences include limitations in localization accuracy, the temporal nature of sound versus the spatial persistent nature of visual stimulation, the limitations on the number of parallel information streams that can be processed by our auditory system due to the phenomenon of masking and the ambiguity in the judgement of perceptual qualities of sounds such as pitch, loudness, level, reverberation and so forth. The development of successful direct manipulation spatial audio systems obviously requires further research. The major research issues include:

- the adequacy of auditory localization as a feedback channel to support aimed movement tasks
- the development of symbolic auditory information and its ability to support metaphors
- the adequacy of auditory symbolic information in supporting complex interaction tasks
- the adequacy of auditory feedback as an indicator of the system and elements state in a complex audio environment
- the formulation of evaluation methodologies for objective comparisons on the aforementioned issues

Of the aforementioned issues the only one where there is a body of research is the development of symbolic information for auditory display. The thesis will focus on the first two of these issues.

3.8.2 Auditory Signs

The formal study of signs requires a classification however the plethora of signs can easily lead to a rather daunting list. For this reason, a division based on the level of arbitrariness of each sign is commonly used that uses three main categories [26]. According to this division a sign can be:

- Symbolic, denoting a fundamentally arbitrary relationship between signifier and signified that has to be learned
- Iconic, denoting a relationship based on the resemblance of the signifier to the signified that can be directly perceived
- Indexical, denoting a relationship that can be denoted by the signifier through a inference process

Gaver [43] tried to transfer the notion of iconic conventions for sign construction into the audio domain. Gaver's terminology is slightly different from one used in semiotic literature and he refers to signs as perceptual mappings, used to enable us to perceive the model world of the computer (the metaphor used to present the computer). He classified perceptual mappings into symbolic, nomic (Iconic) and metaphorical which as can be seen in Table 2, is a rephrase of the terminology used by semiotics. Blattner [13] also proposed certain methodologies to create auditory signs. Although not directly referring to semiotics (Blattner refers to Marcus [79]), he also subdivides sign mappings into representational, abstract, semi-abstract. Blattner does not refer directly to the notion of signs, he thinks of all these sub-

categories as icons. Blattner's icons are essentially what semiotics refer to as signs, a fact that essentially reveals the analogy between semiotics and Blattner's and Gaver's definitions.

Sign Divisions		
Gaver	Blattner	Semiotics
Symbolic	Abstract	Symbolic
Metaphorical	Semi-Abstract	Indexical
Iconic	Representational	Iconic

Table 2. A summary of views on sign divisions and the associated terminology.

In looking for signs for auditory display, researchers utilized the abilities of human auditory perception. From a psychological point of view, auditory pattern recognition enables humans to perform three listening modes: everyday listening, musical listening and speech processing. Musical listening refers to attributes of sound (pitch, timbre, rhythm, dynamics) and their interplay per se, whereas everyday listening refers to the process of listening for the hearing attributes of the sound generating events. To give an example, the process of understanding that it was a wooden door that just closed behind us is one of everyday listening while the process of observing the pitch interplay between two musical instruments is one of musical listening. Speech processing refers to the particular process of auditory pattern recognition that enables us to perceive and process speech.

It should be mentioned that the distinction between listening modes is not one of the physical auditory event; it rather refers to the listening experience and is therefore of psychological nature. Signs for audio displays are derived from the pool of our pattern recognition potential and can be classified as speech sounds, auditory icons [43] and earcons [13].

Sound design based on auditory icons is a methodology that builds on everyday listening and uses iconic and indexical mappings for the creation of auditory signs. Objects in the display are assigned physical properties, like material, size etc and when interaction with the objects is performed a corresponding sound is played. The methodology was created to supplement visual displays with sound, covering an important design gap of that period.

Earcons provide an alternative design methodology based on structured musical listening. The design elements for earcon design are sound attributes as these are experienced through musical listening, timbre, pitch, rhythm, register and dynamics. The aforementioned parameters are manipulated for creating earcons. There are two main earcon categories, one-element and compound. One-element earcons may be digitized sounds, a sound created by a synthesizer, a single note or a motive. One-element earcons may be used to represent simple, basic or commonly occurring user-interface entities such as key-clicks, cursors, or selection mechanisms, common error messages etc. In this way a primary set of earcons is created that can be used consequently to create compound earcons. Compound earcons are essentially

used to provide feedback for more complex display operations. Three construction methods exist for creating compound earcons, combining, inheriting and transforming. Combined earcons are created by placing two otherwise existing earcons in succession. Using inheritance operations hierarchical earcon families can be derived. Blattner proposes to use un-pitched rhythm for the most abstract level, followed by pitch, timbre and finally register and dynamics for the different levels of hierarchy. Earcons have been extensively evaluated by Brewster [17, 20]. McGookin [80] proposed using spatial audio to improve the comprehension of simultaneously presented earcons.

Sound design based on speech, is a methodology that essentially uses symbolic, linguistic mappings. Such a design methodology is very powerful since any type of meaning can be presented. Speech based interfaces have been successfully used in interfaces for blind users, for a thorough review the reader is directed to the book by Raman [99]. The rationale for using speech in an audio display is essentially the same as using text in a visual display. Speech is useful as a supplementary tool when other types of auditory signs do not suffice and is essential for the presentation of documents containing linguistic information.

Overusing speech will however result in an annoying display. It is worth noting that the time it takes for a spoken message to be said increases with the length of the message, therefore over-utilization of speech in an interface may result in slow information presentation. Iconic and Indexical auditory signs however, are quicker to present and if successfully designed faster to perceive and interact with. In this way, they can carry information in a more efficient way and for this reason they can be used for building the display core and basic metaphor. However, their interpretation can be ambiguous and therefore their usefulness might be compromised in situations where clarity is of critical importance.

3.9 Discussion

The review provided in the beginning of this chapter reveals the state of the art in spatial audio display research. Judging from the number of application areas that spatial audio has been applied it is possible to conclude that although the field is new it has great potential. Spatial audio could play an important role in fields such as users with visual impairment, virtual reality, mobile computing and as a feedback channel for visual systems. Certain metaphors such as time to space and the metaphor of a spatially positioned speaker speaking out documents have been found to be supported by this modality.

It is however, evident that the development so far has been based on the intuition of the developers and few research results into the design of spatial audio displays are available. These findings can be summarized in that interaction with an egocentric display is faster compared to that in an exocentric, that spatialised audio assists the perception of simultaneous speech streams in teleconference settings and in that exocentric spatial audio can be used to assist navigation tasks, however increased homing times are commonly observed. These findings also verified that spatial audio could be successfully used to convey background information for users engaged in a primary visual task. It can also support certain metaphors

such as the one of time to space, where certain display locations correspond to time attributes of the object of interest, such as time of arrival.

The design of spatial audio displays has been based so far mostly on the direct manipulation paradigm. Designs have been either egocentric or exocentric. There is however, a substantial gap in evaluation studies that show how to design such systems in a usable way. In all cases, sounds are placed in the frontal horizontal plane with a few exceptions that use front back cues to differentiate between the front and the back of a user. Such a choice is justified from a psychoacoustic point of view since defining the display to be in that area alleviates by design the problem of confusions. Indeed that large confusion rate observed when using non-individualized HRTFs is a major obstacle in expanding the display area. It would be interesting in this sense to investigate whether the display area could be extended by using individualized HRTF functions which reduce the confusion rates users would experience.

The pointing based interactions that are encountered in direct manipulation displays have not been considered in a fashion similar to the one encountered in visual displays as this is found to be the case in Chapter 4. Only Friedlander [39] investigating pointing to audio targets in a matter similar to the one used in visual pointing tasks, however in this study the targets were not spatialised and there was no direct contact with the display elements. In that sense, the issue of spatial audio target acquisition has not been examined. No modelling effort has been tried and in addition the effect of target width and distance to target in spatial audio displays remains to a great extent unknown. Investigating such issues can provide insight on issues relating to display segmentation and scalability. Most of the studies so far have been using up to four sources however their alignment and presentation mode has not been investigated in detail. In that sense questions, like how many items a display can accommodate and how they can be presented cannot be answered.

Sound design for spatial audio displays is also unspecified to a large extent. Although the concepts of auditory icons earcons and speech are relevant, their ability to support metaphors and enable interaction in a purely audio environment has not been examined. Finally, the different gestures that have been used in the literature have not been tested against each other and it is unclear in what situations they are more appropriate.

In this sense, designers are left to their own devices and cannot proceed based on guidelines and design recommendation as is the case in vision based applications. This is largely due to the fact that most of the design proposals have not been evaluated and therefore the shortcomings of the designs in the literature have not been revealed. The thesis aims to cover this gap by evaluating the fundamental task of spatial audio target acquisition.

Issues related to display presentation, the creation of metaphors based on semiotic audio information and of interaction in more complex display arrangements are not treated in thesis. In the rest of the text spatial audio target acquisition is examined and based on the findings a selection task and display segmentation guidelines are obtained. Such information can be used in the future to form the basis

of more complex display arrangements where issues related to presentation and interaction in more complex displays can be examined in detail.

3.10 Conclusions

This chapter reviewed the design trends in spatial audio display research. Display designs, reproduction options, control and possible application areas have been presented. The review indicates that, although a substantial number of proposals exist for different application areas, the literature is short of studies focusing on the design aspects of such displays. The thesis subject of disambiguating spatial audio target acquisition in the manner presented in the Research Questions is found to be novel and to contribute to the already available body of research in the field.

The material developed in this chapter reveals that the creation of successful direct manipulation spatial audio interfaces requires the verification of the ability of our hearing sense to provide:

- Adequate differentiation of display elements in a spatial sense
- Support for control based on aimed movements
- Creation of successful semiotic mappings to communicate a valid conceptual model
- Support for closed loop control based on audio feedback to help users to evaluate the results of their actions and validate their conceptual models

Of these questions none apart from the third one has been so far evaluated in the context that has been defined by contemporary Human Computer Interaction research. In order to achieve the creation of successful direct manipulation spatial audio interfaces it is necessary to perform studies that will shed light on the appropriate design options for supporting it. The thesis focuses on whether our auditory sense can provide adequate support for control based on aimed movements. The type of studies necessary for the disambiguation of the aforementioned research issue are controlled user studies that aim to evaluate the usability of spatial audio target acquisition as an interaction method in direct manipulation interfaces. To draw inspiration for the design and the conduction of such studies, the thesis looks into the existing body of research that has been developed in evaluating aimed movements in the visual domain (Chapter 4) as well as psychoacoustical research to understand the potential of directional hearing. This information is considered essential for answering the research questions set by the thesis which are repeated here.

RQ 1 How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?

RQ 2 What are the factors that affect deictic spatial audio target acquisition?

RQ 3 How can we evaluate the usability of deictic spatial audio target acquisition?

The next chapter is providing a review of existing methodologies for the evaluation of aimed movements.

4 Evaluating & Modelling Pointing Interactions

4.1 Introduction

Understanding human aimed movement is critical in the context of the thesis because it will provide insight into how to create a usable spatial audio selection task. Given that visual aimed movements have been found to be successful in the context of contemporary human computer interaction systems, understanding how they are performed, will improve our understanding of the enhancements required in the design of spatial audio target acquisition tasks.

Evaluating, modelling and understanding motor behaviour in humans reaching for visual targets using a part of their body is an active research area in psychology, ergonomics and human computer interaction. In psychology, research is focused in modelling and understanding human motor behaviour in what has been called *aimed movements*. Such research has had great impact in ergonomics and human factors, since the psychophysical models shed light on the effectiveness of different control mechanisms and presentation modes used in physical or virtual interactive systems. A number of competing models of performance in *aimed movement* or otherwise stated *target acquisition tasks* have been proposed. These models can be classified into three main categories: logarithmic, linear and power models. They focus mainly on modelling the time required to reach a target location as a function of the *amplitude of movement* (an alternative term for distance to target) and *target width*. The application of the models has been found to be valid for movements performed by a variety of parts of the human body, such as hand, fingers, head and wrist movements.

In addition, the behaviour predicted by the models for natural human movement has been found to be valid for modelling the control of virtual pointers, a fact that makes such studies especially relevant to human computer interaction. Two main theories have been developed to account for the descriptive power of the models: the deterministic and the stochastic iterative corrections theories [82]. The theories differentiate between an initial ballistic submovement towards the target followed by corrective submovements that help direct the movement into the target using visual feedback. This chapter presents existing models, the theories developed to explain the rationale for the success of the models and a number of relevant studies that illustrate the application of the models and their predictive power.

4.2 Empirically Derived Models

Models of human movement are mainly empirically derived, based on observations of time patterns during aimed human movement studies. The most common experimental setup for aimed movement studies is the one presented in Figure 14. Participants are asked to reciprocally move to one of the two targets and movement amplitude and target width are controlled by the experimenter. Dependent variables are time, error rates and more recently, movement displacement and velocity patterns. Quite

commonly target width and distance to target are adjusted in between blocks of trials and participants are subsequently asked to repeat the task. Accuracy of movement has been considered a side factor until recently, due to the fact that target width in most of the studies is such that low error ratios are usually expected and observed. Research is focused in observing movement time under normal conditions. Net movement time is considered when possible by the experimental design. Time measurements should therefore not include the time to initiate movement (reaction time) or the time to indicate movement termination. However, in some cases limitations in measurement equipment and experimental design cannot support such an option. All of the studies model movement time as a function of the ratio of distance to target to target width (A/W). Movement displacement, velocity and acceleration are also considered by recent studies as these are registered by tracking technologies, such as inertial, electromagnetic or camera based sensing.

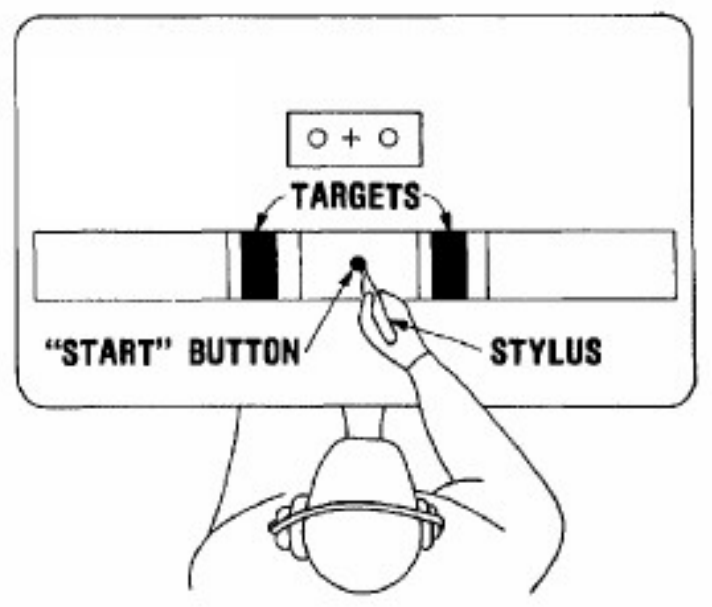


Figure 14. Fitts' Experimental Design, a participant selecting two targets in a discrete fashion. Participant is shown resting in the home position waiting for a signal to start moving to the indicated direction (adapted from [37]). Participants return to the home position after selection. When participants alternate between the targets continuously, the task is called continuous.

The experiments investigating aimed movement in psychology require participants to perform the task in either a spatially or a temporally constrained manner. This is ensured by the experimental instructions which may ask participants to concentrate in finishing their movement inside the target region or to concentrate in performing their movement within a certain (usually short) time frame. In the heart of the manipulation is the so called '*speed-accuracy trade off*' that is inherent in aimed movement. The trade off essentially describes the intuitive observation that more time to move to a target will result

in greater selection accuracy. Spatially constrained tasks mainly reveal the fine structure of the corrective submovements that occur when participants try to centre their limbs with respect to the target region. As a result, the number of errors is minimal but movement time variation might be considerable. On the other hand, temporal constraints reveal more of the characteristics of the initial movement towards the target area. Consequently, error rates and movement endpoint variability are increased, however movement time is better controlled. In human computer interaction literature the applied nature of the research results in experimental instructions where participants are required to focus both on the speed as well as in the accuracy of movement.

It is useful to distinguish between two styles of such studies: those facilitating discreet and those facilitating continuous movements. In the continuous case, participants are instructed to continuously alternate between the two targets. In the discrete case, participants are instructed to wait at a home position, usually at the centre of the experimental platform, until a signal that indicates the movement direction for the next trial is presented and subsequently start to move. After movement is finished participants are asked to return to the home position and wait for the next trial.

The aforementioned experimental setups are typical of studies involving real targets and physical movement. In recent studies, both in psychology as well as in human computer interaction and ergonomics, virtual targets presented on a computer screen are commonly used. In such cases, participants perform the aimed movements using virtual pointers controlled by hardware devices such as mice, joysticks etc. In such studies, targets can be presented in different fashions such as in a cyclical pattern, where manipulations of more independent variables such as angle of approach to the target, is feasible. Despite the alterations of the experimental paradigm obtained data are consistent with data from physical movement studies, in that way verifying the global nature and validity of the models that were developed using the early experimental setups.

4.2.1 Effective Target Width

Considerable variability with respect to aimed movement endpoints has been observed in aimed movement studies. This variability provides a well founded indication that noise is inherent in aimed movements. It has been shown that movement endpoints follow a normal distribution around the centre of the target area [36]. This leads to the conclusion that human movement is perturbed by white noise. For target widths that allow sufficiently easy targeting, this movement variability results in certain error rates. Experimentation however has shown error rates to depend on the Index of Difficulty [31, 82, 107], increasing with increasing target amplitude. Wade, Newell and Wallace as well as Card *et al.* [24, 120] have also found a significant main effect between error rate and target width, with errors increasing as target width decreased but no main effect between error rate and distance to target. The range of errors in such cases is around 4% [74], a criterion that can be used to infer the optimality of a target width choice.

For very narrow or too wide targets, observations reveal that movement endpoints can be concentrated in areas that might not correspond to the experimentally defined target area and can be either smaller or larger. This situation occurs when participants act as if relatively narrow targets are wider than what they really are or as if relatively wide targets are narrower than they really are. This in effect can compress the range of subjective target widths compared to the range of objective target widths making changes in objective target width have less effect than they would have otherwise.

These observations led to the introduction of the quantity of effective target width [74, 126]. Effective target width essentially represents the human perspective of the target areas as this appears by observing the histograms of the endpoints of aimed movements. This subjective nature of the effective target width concept led some researchers to use the term subjective target width instead of effective target width, as in [82].

The rationale behind the introduction of effective target width is to improve the goodness of fit of aimed movement time models. However, effective target width has not been used extensively in Fitts' law studies with the exception of [74, 82, 126]. In fact, MacKenzie [74] found that using effective target width instead of objective target width during model formulation did not prove to provide significantly better results.

Apart from improving the goodness of fit, effective target width can be used as an indicator of the optimality of target width selection. The rationale for such an application is that when effective target width is less than objective, the target is too large for a user to select. In the opposite case it provides an indication that a target is too small to select. In this way, target size for certain interaction procedures can be predicted based on the standard deviation of end movements. Such an option, has not been investigated in the literature.

4.2.1.1 Effective Target Width Estimation

In practice, the considerations arising from movement endpoint variability have led researchers to suggest ways to normalize target width (W) to the effective target width (W_e) [75, 126] for modelling purposes. MacKenzie [74] citing Welford [126] presents two methods for performing the normalization. Effective target width can be calculated as in Equation **1**based on σ, the standard deviation of the movement endpoint.

$$W_e = 4.133\sigma \qquad \textbf{Equation 1}$$

The derivation is based on the fact that for a normal distribution of variance σ, an area of size W_e would contain 96% of the observations (endpoints). An error rate of 4% is therefore expected, even under optimal aimed movement conditions. Such an error rate indicates that subjective and objective target widths coincide. Error rates smaller or higher than 4% provide an indication that subjective and objective

target widths do not coincide. In such cases, if the error rate is known, say n%, it is possible to calculate the effective target width using a table of z scores. The procedure starts by determining a z value such that ±z contains 100-n percent of the area under the unit normal curve. Effective target width is given in this case by

$$W_e = 2.066/z \times W \qquad \textbf{Equation 2}$$

4.2.2 Logarithmic Models

The very first modelling effort was the one by Fitts [37]. Fitts based his modelling approach on the information theory modelling perspective that is reviewed in Section 4.3. The modelling effort is developed around the central concept of the Index of Performance. According to Fitts [37] the index of performance can be calculated as in Equation 3:

$$I_p = \frac{\log_2(2A/W)}{MT} \qquad \textbf{Equation 3}$$

In Equation 3 W is the target width and A is the amplitude of movement to the centre of the target area. MT is the movement time required to reach the target. The logarithmic term is called the *Index of Difficulty*. The Equation effectively states that movement time is proportional to the Index of Difficulty and inversely proportional to the Index of Performance. This statement predicts a linear relationship between movement time and index of difficulty. Equation **4** provides the model.

$$MT = a + b\log_2(2A/W) \qquad \textbf{Equation 4}$$

The parameter *α* (also called *intercept*) and parameter b (also called *slope*) are task dependent parameters estimated using linear regression. If parameter α is close to zero, the inverse of b is an estimator of the Index of Performance. Time to move to a target has been found to correlate with Equation **4** see for example MacKenzie [74]. It has been argued that parameter α when net movement time is considered, should equal or be very close to zero. However, this is rarely the case. Non-zero values are mostly attributed to additive factors affecting movement initiation and termination time. Movement termination time is an artefact of the experimental design. It corresponds to the time required to terminate a trial, in many such computer related studies, it is associated with the time required to click the mouse button, a common method used to terminate a trial. Movement initiation time sometimes cannot be eliminated due to measurement limitations in the experimental design.

There is some controversy in the literature on logarithmic models mostly with respect to the definition of the index of difficulty. Welford [127] proposed the following model, which has been found

to give a better fit to data compared to the Fitts' formulation for low indices of difficulty as mentioned in [75] .

$$MT = a + b\log_2(\frac{A}{W} + 0.5)$$ **Equation 5**

MacKenzie and Buxton [74, 76] proposed another variation where target width is associated with movement noise and the signal term to the amplitude of movement to result in Equation 6. MacKenzie's formulation is the most contemporary alteration proposed to Fitts' law. It has been found [76] to provide a better fit to experimental data, compared to Fitts' formulation.

$$MT = a + b\log_2(\frac{A}{W} + 1)$$ **Equation 6**

4.2.2.1 Benefits & Disadvantages of the Logarithmic Models

Fitts' law has proven to be a very useful tool for designers. No matter what problems appeared with the model, it has always been found to correlate well under 'normal' conditions. In addition, it inspired a lot of research in human motor behaviour and it has been used in organizing workspaces in environments where repeated movements are necessary. In addition, the rationale behind the Index of Performance inspired experimental manipulations that served to characterize the efficiency of different movement tasks and interaction techniques. The idea is to calculate the Index of Performance, for a number of devices, under the same experimental task and then used this quantity to characterize the efficiency of the devices.

One of the issues that restrict the application of Fitts' law is its one dimensional nature. In human computer interaction contexts targets are essentially two dimensional and therefore there is a certain ambiguity with how the law must be applied. In such contexts, the formulation of the index of difficulty as it appears in Equation 4 and Equation 5 can lead to negative values, if A becomes less than W/2. A negative index of difficulty does not have a physical interpretation. For one dimensional selection tasks such a prediction is not very unreasonable since when the movement amplitude required to reach the target is less than W/2 the movement starting position is already inside the target area but for two dimensional targets this is a serious problem. The issue is examined in detail in Section 4.4.1.

MacKenzie's formulation is the only one providing a positive value for the index of difficulty for all possible A and W values. The formulation has been proven useful in applying Fitts' law in two dimensional target acquisition and it should be noted that it affords better modelling of additive factors.

Fitts' formulation has also been criticised for lack of fit in the region of small indices of difficulty, ID values around or less than 2. In Fitts' original experiment there was considerable variation for the

index of performance when the index of difficulty was 2. For indices of difficulty between 3 and 7, Index of Performance varied between 9.58 and 11.54 bits per second, however it was only 5 for Index of Difficulty of 2, data available by Fitts as well as MacKenzie [37, 74]. In many studies, a lack of fit of the law for low indices of difficulty has been observed. Trying to follow this data trend revisions of the original model were proposed in Equation 5 and Equation 6. Equation 6 results at smoother slopes at lower Indices of Difficulty, a fact accommodated by the additive terms in the logarithm with values of 0.5 and 1 respectively. This affordance explains to some extent the better fit of these equations to data especially to lower indices of difficulty. However, all equations predict that the smaller the amplitude of movement the less time it should take to perform. This might not be exactly so considering that the noise fluctuations in the human motor system result in lower signal to noise ratios for relatively short movement amplitudes (weak signals) compared to long movements.

Furthermore, Fitts' formulation of the index of difficulty suggests that distance to target is influencing time to home twice as much as target width. This argument does not appear to hold under careful examination. Sheridan [108] showed that reductions in target width cause a disproportionate increase in movement time when compared with similar increases in target amplitude. This has also been independently verified by Meyer *et al* [82]. Equation 5 and Equation 6 are improving Fitts' law in this aspect, predicting a contribution of the same magnitude in movement time for the quantities of distance to target and target width.

Finally, all logarithmic models have been developed for spatially constrained or dual tasks, in both cases tasks where accuracy of movement endpoint was important and corrective submovements were taken into account. It has been reported in the literature [82, 107] that they cannot account as satisfactorily as other model types for temporally constrained tasks.

4.2.3 Linear Models

Linear models have been found to apply in two notably different situations. The first is temporarily constrained tasks and the second is in movements through constrained paths.

With respect to the first case the first observation for a linear model based on movement time data from a temporally constrained task has been done by Schmidt *et al.* [107]. As has been mentioned in Section 4.2, the characteristic of temporally constrained tasks is the relatively higher movement endpoint variability that is the result of the relatively strict time constraints. According to Schmidt's '*impulse variability model*' the standard deviation in endpoint coordinates was found to be a linear function of average velocity calculated as distance over time as in Equation 7. In Equation 7 A is the amplitude of movement and MT the movement time.

$$W_e = a + b\,A/MT \qquad \textbf{Equation 7}$$

The model essentially predicts a linear relationship between movement time and A/W for temporally constrained tasks. Schmidt [107] verified these predictions for movements over relatively

short distances. However, the relationship was not verified for movements that lasted longer than 200 ms. 290 ms is considered the time limit user which the movement is ballistic and controlled mainly by kinaesthetic feedback. After this time period external feedback through the senses is used, a factor that makes movement time a logarithmic function of the index of difficulty. An intermediate situation where the task is temporarily constrained but a secondary submovement is allowed is related to the power model presented by Meyer *et al* [82] and is presented in Section 4.2.4.

A linear model on the time to complete a movement under substantially different conditions has been proposed by Friedlander *et al* [39]. In this study, homing to non-visual targets in a bulls eye (marking) menu based on auditory or tactile feedback was investigated. In each trial of their experiment, participants were asked to move into one out of four main menu directions while counting certain steps that were associated with the ring width in the bulls-eye menu. Audio and tactile feedback had been tested as a means of marking ring widths to define the target in the display. According to the results, a linear model was more appropriately accounting for movement times. This implies that participants followed a different behaviour in their targeting strategy. The authors verified that distance to target and target width indeed affect time to select. The formula that the authors suggest for the approximation of time to target is

$$MT = \alpha + \beta \cdot A/W \qquad \textbf{Equation 8}$$

In the Friedlander *et al* study participants were only informed on the direction they should move and the number of concentric circles they had to cross to reach the target. The fundamental difference between the two tasks is that in Fitts' law studies participants are perceptually aware through their senses of target position, whereas in the Friedlander case participants verify they have reached the target through a cognitive process. It is interesting to observe that A/W is the number of concentric circles participants had to cross, a fact revealing movement time in this case was proportional to the number of items participants had to cross to get to the target.

Although not directly related to aimed movement studies it is interesting to refer to a model of Accot & Zhai [1], namely the steering law which conforms to Equation **8**. The steering law applies to trajectory based tasks where the user has to travel along a constrained path. Such a trajectory is depicted in Figure 15.

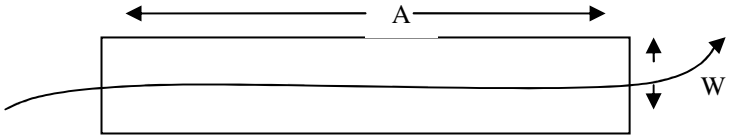


Figure 15. Illustration of the trajectory encountered when a user moves the a virtual pointer through a constrained path, of length A and width W. A linear model has been found to predict movement time in such a case.

4.2.4 Power Models

The general form of the power models is given in Equation 9 as proposed by Kvaolseth [64]. In Equation 9, A and W are distance to target and target width and a, b, c are constants estimated by regression.

$$MT = aA^{b}W^{c} \qquad \textbf{Equation 9}$$

Equation 9 has been found to give better fit compared to the logarithmic or linear models. However, this is greatly attributed to the extra degree of freedom in the equation. If b = -c then it takes the form of

$$MT = a\left(A/W\right)^{b} \qquad \textbf{Equation 10}$$

If the base 2 logarithm is taken, Equation 10 can be used to predict the base 2 logarithm of movement time in a manner similar to Fitts' law. Meyer *et al*. [82] proposed a power model for predicting movement time where b=1/2.

$$MT = a + b\sqrt{A/W} \qquad \textbf{Equation 11}$$

The model in Equation 11 can be derived analytically based on certain theoretical assumptions with respect to user movement pattern and movement endpoint variability. Section 4.3 refers to the rationale behind the model. In short, Meyer and colleagues tried to provide a unified conceptual framework encompassing both the linear as well as the logarithmic model, in that way trying to describe performance combining results from spatially and temporarily constrained movement tasks. They verified that visual target acquisition consists of a primary followed by corrective submovements. In the primary submovement a linear speed/accuracy trade off is observed whereas the corrective submovements result in a logarithmic speed accuracy trade off. Based on this experimentally verified hypothesis on submovement nature they provided an analytic derivation for total movement time that resulted in Equation 10 which was verified to account well for total movement times.

4.3 Theoretical Investigations & the Speed-Accuracy Trade-Off

Based on the review of models of aimed movement to visual targets, it can be observed that a number of ways exist to model aimed movement to visual targets. It is within the scope of the thesis to identify which of the models is most appropriate for the task under examination. So far, it can be argued that a factor that will influence the applicability of each model is the nature of the task, i.e. whether the task is spatially or temporarily constrained. However, substantial information can be obtained by also

examining the theory behind each of the models. This information is mostly related to how people perform these movements and what kind of feedback they use in the process. The next section presents the theory behind each of the models and tries to identify components of the theories that can be used to explain the acquisition task that is going to be examined in the thesis. Apart from examining the aforementioned issue theoretically experimental investigation is provided in Chapters 7 and 8.

4.3.1 The information theoretic perspective

Logarithmic models stem from a research school, mainly in psychology, that claims that aimed human movement can be modelled based on information theory. Human performance theories of this school seek an approach to the study of information processing within the nervous system and in relation to communication with the environment. This approach was applied in a number of studies on different information processing related topics, such as choice reaction time. In such a task it has been observed that reaction time increases with the logarithm of the possible options and linearly with the amount of information in the stimuli. Of critical relevance to aimed movement studies was the calculation of the information capacity of the human motor system. The informational capacity essentially represents the rate at which the motor system transmits information. This cannot be considered constant and it depends on the amount of information transmitted, training and task characteristics [36].

Fitts proposed a method for calculating the informational capacity of the motor system as the ratio of the difficulty of a certain task expressed in bits, divided by the time required to complete the task. The difficulty of the task was seen by Fitts as related to the information inherent in the task. He therefore used the base-2 logarithm of Weber's ratio to express the difficulty of movement task of amplitude A to a target of width W. If seen from the Weber's law point of view, the amplitude of movement corresponds to the stimulus and W/2 to correspond to the noticeable difference with respect to the stimulus. According to Weber-Fechner's law [35] the differential change in perceived magnitude p is a function of the differential increase in the stimulus dS and the value of the stimulus S as given by Equation 12.

Weber's law assumes the constant k to be one however, it has been subsequently found that in certain situations this is not the case.

$$dp = k\frac{dS}{S} \qquad \textbf{Equation 12}$$

If Equation 12 is integrated it results in Equation 13, where S_0 is the threshold of the stimulus below which it is not perceived at all.

$$p = k\ln S - k\ln S_0 \qquad \textbf{Equation 13}$$

Applying the law in the perceived difficulty of the movement Fitts' set S to be the amplitude of movement and S_0 corresponds to half the target width, the area out of width the movement is not considered valid anymore. Substituting in Equation 13 results in the familiar expression of the index of difficulty as this is described by Fitts. It is given again here in Equation 14.

$$ID = \log_2 2A/W \qquad \textbf{Equation 14}$$

Fitts used the base 2 logarithm in the integration instead of the natural logarithm so that the result can be calculated in bits in accordance to the information theory. The capacity of the system is subsequently calculated by dividing ID to the movement time as in Equation 1. It is therefore evident that Fitts applied the information theoretic perspective into the difficulty of a task as this has been estimated using Weber's law.

A different approach that results in a logarithmic model is the one presented by MacKenzie and was provided in Equation 6. MacKenzie provided a model that makes direct use of Shannon's theorem 17 quoted in [37, 74], which states that the effective information capacity C of a communications system of bandwidth B expressed in bits/sec as in Equation 15.

$$C = B\log_2\left(\frac{S+N}{N}\right) \qquad \textbf{Equation 15}$$

In Equation 15, S is the signal power and N the noise power. Assuming that the signal corresponds to the amplitude of movement (A) and that noise is represented by the width of the target (W), MacKenzie proposed Equation 6. This equation is characterised by the fact that it conforms purely with the information theoretic point of view without making use of Weber's law to estimate perceived difficulty. In practice the equation is not markedly different than the one proposed by Fitts except for the difference that the rate of decrease in movement time is smoother for low ID levels a fact that conforms to empirical observations. In addition, it provides an equal weight in the determination of movement time for amplitude to target and target width fact that conforms better to empirical results.

4.3.2 The iterative corrections model

An alternative way to explain Equation 14 was provided by Crossman and Goodeve [31]. The model assumes that an overall movement from an initial home position to a target region includes a series of discrete submovements made on the basis of sensory feedback. Each submovement supposedly has a well defined beginning and an end, takes a constant time increment to complete (t) and travels a constant proportion (p_d) of the remaining distance to the centre of the target. The model is deterministic in the sense that it incorporates no random variability due to neuromotor noise or other stochastic factors. The extents and the duration of component submovements are assumed to have fixed values across movement

trials involving the same combination of target distance and width. Termination of submovement series occurs as soon as a submovement ends inside the target region. According to the model at submovement *n*, the total distance covered is

$$p_d A(1-p_d)^{n-1} \qquad \textbf{Equation 16}$$

In Equation **16** A is the amplitude of movement required to reach the target. The remaining distance after the nth movement is

$$p_d^n \qquad \textbf{Equation 17}$$

Solving for n yields

$$n = -\frac{1}{\log_2 p}\log_2 \frac{2D}{W} \qquad \textbf{Equation 18}$$

According to the model, each submovement takes a total time t, so the total time to the target is *n·t*, or

$$MT = C\log_2\left(\frac{2D}{W}\right) \qquad \textbf{Equation 19}$$

C is a positive constant determined by

$$-\frac{t}{\log_2 p} \qquad \textbf{Equation 20}$$

The logarithm base two is used when solving for n in order to provide consistency with the information theory formulations that are expressed in bits. One interesting prediction of the model is based on the idea that each ballistic submovement is organized so that the error can be reduced on the basis of feedback from the preceding submovement. Accordingly, a higher number of corrective movements will improve the accuracy of aiming but it will increase the total movement time. This observation is the foundation of what has been termed *'speed-accuracy trade off in goal orientated aiming'*.

4.3.3 The stochastic optimized submovements model

The deterministic iterative corrections model prevailed for a period of about twenty years. A significant contribution of the model is the realization that a series of movements is required to perform

aimed movement. A primary submovement is used to bring the limb close to the target and secondary submovements to fine tune the accuracy of the movement. An illustration is provided in Figure 16.

However, recently obtained kinematical data show that the model has to be refined. In particular, there is considerable variability to submovement time and final positioning dependent on distance to target and target width, according to experimental data mentioned in [82]. In addition, when the action is performed in visual feedback deprivation conditions as in [134], the variability of the primary submovement is particularly pronounced. This variability is not accounted by the model due to its deterministic nature [82].

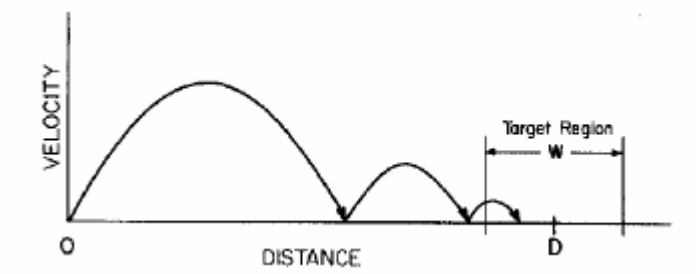


Figure 16. Outline of the assumptions of the deterministic iterative-correction model with respect to movement trajectory. The horizontal axis represents distance and the vertical axis velocity. The curves correspond to successive submovements between an initial home position and a target region.

A rather wide spread view in the literature is that the number of submovements is dependent on the time available to complete the task. Experimentation [11, 25] showed that for time constrained movements in the period of less than 290ms such an option is not possible. Schmidt et *al*. [107] studied in detail the effect of time constraints in the movement. They found that for short, time constrained movements the variability in the muscular forces results in great variability in the spatial endpoint of the movement. A similar hypothesis for targets in long distances was not verified however, presumably, due to the fact that long distances provide an opportunity for an online correcting process to operate.

These findings led Meyer and colleagues [82] to provide the most contemporary of models namely, the '*stochastic optimized submovement model*'. The model distinguishes between a primary submovement and a secondary submovement that is used to correct the first one as happening in the single correction model. The model differs from the deterministic one since it takes into account the variability that is present in submovement times and submovement endpoints and relates them to factors imposed by the movement context such as movement velocity in a stochastic manner. According to the model the primary and secondary submovement endpoints are assumed to follow a normal distribution around the centre of the target. The standard deviation is assumed to be proportional to movement

velocity according to Equation **21**. In Equation **21** D_1 is mean displacement of the primary submovement, T_1 is the mean primary submovement time and K is a constant.

$$S_1 = KV_1 = K\,D_1/T_1 \qquad \textbf{Equation 21}$$

If Δ is the deviation of the primary submovement endpoint from the target, then if

$$|\Delta| \geq W/2 \qquad \textbf{Equation 22}$$

a second submovement follows whose endpoints standard deviations are according to Equation **23**. In Equation **21**, $T_{2\Delta}$ is the mean time of the second submovements Δ is the deviation of the primary submovement endpoint and K is a constant.

$$S_2 = K\,\Delta/T_{2\Delta} \qquad \textbf{Equation 23}$$

Based on the aforementioned models and rather complex algebraic manipulations, Meyer *et al*, provided equations that predict the time and the error rate of aimed movement. It is worth noting that this study was the first that considered error rate explicitly as a function of A/W. Movement time is calculated of the some of the first and the second submovements. Meyer *et al*. considered only two submovements; however this is a hypothesis that has not been verified. Furthermore, for subsequent submovements the equations become prohibitively and unnecessarily complex, considering especially the simplicity of Fitts' law and its high correlation with different kinds of movement data.

$$MT = 2K\left(2\theta\sqrt{D/W} - \sqrt{W/D}\right)/\,\theta\sqrt{\theta - (W/D)} \qquad \textbf{Equation 24}$$

Where K is the slope of the speed accuracy trade-off and θ is a parameter calculated by an iterative procedure. Error rates are given by

$$p_e = (1-c_1)(1-c_2) = 2(1-c_2)\left(1 - N\left[1/\sqrt{\theta(D/W)-1}\right]\right) \qquad \textbf{Equation 25}$$

In Equation 25 c1 and c2 are the corresponding probabilities that the primary and secondary submovements end inside the target regions and N(x) denotes the probability that a random variable with zero mean and unit variance is less than x. c2 is assumed to be constant and determined by experimental data, for Meyer's study it was 0.95.

In conclusion, the review indicates that aimed movements to visual targets are performed using an initial ballistic movement followed by corrective submovements. A successful acquisition task should therefore provide feedback to support these movements. In the following section a number of studies that verified the applicability of the aforementioned models are reviewed to provide evidence on the large

scope of models. Such information is useful since it indicates that with successful design spatial audio target acquisition could be modelled in a similar way.

4.4 Applications in Human Computer Interaction

Fitts' law has been extensively used in the context of human computer interaction to model the aimed movement to virtual targets. The number of studies that involve Fitts' law in some way are numerous and a complete review is out of the scope of this chapter. This chapter focuses on a few studies in order to inform the reader on the suitability of Fitts' law and the ways it can be used as a tool that allows useful comparisons that provide insight mainly with respect to the efficiency and effectiveness of interaction.

The first researchers who reported promising results for the application of Fitts' law in human computer interaction were Card *et al.* [24]. They verified that Fitts' law applies in targeting virtual visual targets. This work was also the first to use Fitts' law as a method of comparison between devices. The idea is that when a model for interaction with the device is developed it can subsequently serve for the comparison of the device to other ones. In particular, the quantity of the Index of Performance can help to compare devices. The particular study targeted selection of text characters, using a mouse, an isometric joystick, text keys and step keys. Independent variables were distance to target in cm, target width in characters and approach angle in degrees. The effect of learning was also quantified by asking participants to perform the task in subsequent blocks. The mouse was found to be the most efficient device with IP=10.4 bits/sec, joystick following with 4.5 bits/sec. With respect to movement times, mouse was the fastest (MT = 1660ms), followed by joystick (MT = 1830 ms), followed by text keys (MT = 1830 ms) with last being the step keys (MT=2510 ms). Errors rates averaged from 5% to 13%. Approach angle only had an effect for the joystick increasing time by 3% for approach along the diagonal axis. Welford's variation of Fitts' law was used in this study.

In a study by Drury [32], Fitts' law was used as a tool for foot pedal design. Subjects tapped back and forth between two pedals using their preferred foot. Independent variables were amplitude of movement and pedal sizes. IDs ranged from 0.53 to 2.47 bits and IP was 11.8 bits/sec. Error rate was less than 3.3%.

Jagasinski and Monk [59], applied Fitts' law to a target acquisition task using a displacement joystick and a head mounted control that used two rotating infrared beams. Subjects moved a cursor on a screen to select a circular target using both control options and selected it by staying on target for more than 344 ms. This target selection criterion is interesting since it allows the control of additive factors such as selection time. Experimental factors were amplitude of movement, target size and approach angle in degrees and the two control mechanisms. Indices of Difficulties in the study ranged between 2.0 to 5.6 bits. Movement time and index of difficulty correlated at .99 showing the head control can be a viable

control option. Index of performance was 5 bits/sec. No errors were observed given the special control on the selection criterion.

So *et al*. [115] presented a Fitts' law study where participants were required to select a target using their heads in a head-coupled virtual environment presented through a head mounted display. Participants were controlling a head slave pointer in the virtual environment. Dependent variables were reaction time and movement time and independent variables target width, distance to target and lag between the action of the user and the movement of the virtual cursor. Three lag values were used 0, 133 and 267 ms respectively. Indices of difficulty ranged from 1.32 to 3, with index of performance being 3.8. Lag was found to significantly affect reaction time and also significant interaction was found between lag and target width but not between lag and distance to target. Reaction time was not affected by target width, distance to target or lag.

Fitts' law has also been applied in an eye tracking study by Ware and Mikaelian [125], where targets were selected by three methods, a hardware button, dwell time on target (400 ms) and an on-screen button. Welford's formulation was used, indices of difficulty ranging from -1.0 to 1.8 bits. The highest IP was 13.7 bits/sec. for the hardware button and 9.3 bits/sec for the dwelling time selection procedure. Error rates were quite high ranging from 8.5% for the hardware button to 22% for the on-screen button. As investigators noted eye tracking is a fast selection procedure as long as accuracy demands are minimal.

Another application of Fitts' law with the purpose of comparing three control devices in a benchmark task is the one by MacKenzie [77]. In this study, the same task was performed by participants operating three different devices. By calculating the index of performance MacKenzie was able to conclude on the efficiency of each control device and compare them. The devices tested in the study were a mouse, a trackball and a tablet and the comparison was done both for pointing and for dragging tasks. In this particular study the mouse was found to be as efficient as a tablet in pointing both of them being significantly better than a trackball. With respect to dragging the tablet was found to be the most efficient technique, followed by the mouse and the trackball.

Recent applications of Fitts' law include text entry in mobile phones. Silfverberg *et al*. [114], verified that text entry on phones using either the thumb or the index finger can be sufficiently modelled using Fitts' law. Indices of performance were reported to be 15.625 and 19.2 respectively.

It is worth mentioning again the study by Friedlander et *al*. [39], where targeting in a non-visual bulls eye menu like the one in Figure 17 was investigated. Rings were marked with either tactile or audio feedback and participants were instructed on the direction and number of rings they should cross. As already presented in Section 4.2.3, movement time has found to be a linear relationship of distance to target and target width, however it should be mentioned that Fitts' law also correlated very well with the data.

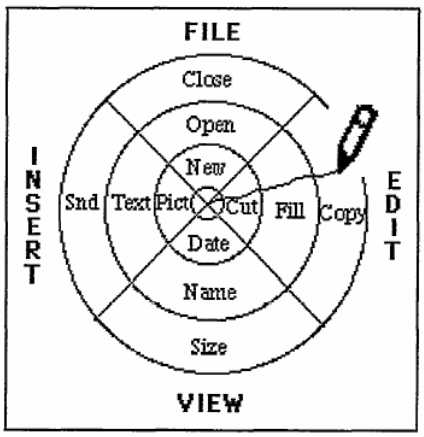


Figure 17. Selection from a bulls eye menu as the one used by Friedlander *et al*. Adapted from [39].

4.4.1 Applying Fitts Law to two and three dimensional targets

A problem that had to be tackled to apply Fitts' law in acquisition tasks in contemporary HCI is the fact that the targets are two dimensional, whereas Fitts' model is one dimensional. However, the analysis involved in such cases is not directly relevant to the thesis since it deals with one dimensional targets. For reasons of completeness, it is just mentioned that MacKenzie [76] examined four candidate models for two dimensional target acquisition. The first substituted width for the projected width value relative to the angle of approach to the target whereas the second called 'the smaller of' model that substituted for W the smaller of either width or height of the target. The third and fourth models have been proposed in a previous study by Gillan et *al.* in [45], the target area and the sum of target dimensions. To accommodate the problem with the negative indices of difficulty Mackenzie proposed the variation to Fitts' that was already presented in Equation 6. According to the results, the 'smaller of' model as well as angle of approach model correlated best to Equation 6 and were considered to associate better with target width when modelling pointing to two dimensional targets. The angle of approach transform is probably the most theoretically solid approach since it can be used as a general method to successfully transform a two dimensional target to the one dimensional quantity required for the application of Fitts' law.

A more recent study on resolving this problem was performed by Accot and Zai [2]. Based on the results of a thorough analysis on a number of competing models and theoretical analysis the authors proposed Equation 26.

$$MT = a + b\log_2\left(\sqrt{\left(\frac{D}{W}\right)^2 + \eta\left(\frac{D}{H}\right)^2} + 1\right)$$ **Equation 26**

In their studies the index of performance was found to be in the order of 5.9 bits/sec. Error rates were in the order of 4%. Grossman and Balakrishnan [51] proposed Equation 27 as a model for trivariate targets.

$$MT = a + b\log_2\left(\sqrt{f(\vartheta)\left(\frac{A}{W}\right)^2 + \left(\frac{A}{H}\right)^2 + h(\vartheta)\left(\frac{A}{D}\right)^2}\right) \quad \textbf{Equation 27}$$

This was verified by the experimental data and index of difficulty was approximately 2. Error rates were in the order of 15%. However, it should be noted that in the experiment small target sizes were used which effectively increased target sizes, however it provided insight on target size for similar interactions.

Finally, Guiard and Beaudouin-Lafon [53] investigated target acquisition in multiscale electronic worlds. They investigated whether Fitts' law would be applicable in the case of a pointing action that involved zooming and panning to reach the target area of a document. The results indicated that Fitts' law could appropriately model movement time for the task under consideration. It is interesting to note that modelling involved an intercept of near to zero value, in accordance with the theoretical expectations of the authors. In addition, the authors showed that view size influences the interaction bandwidth (throughput), however this is only happening in the case of small view sizes. For large view sizes (more than 40 pixels) view size does not influence interaction bandwidth.

4.5 Comments on the models, their scope and their theoretical backgrounds

There is some controversy with respect to the way the models can be applied and which model is better for certain tasks. The most fundamental difference with respect to the model application is the task nature, in particular whether the task is spatially or temporally constrained. In the former case the logarithmic models have been found to be most appropriate. In the latter case, a linear speed accuracy trade off is mostly appropriate for the primary submovement, but when secondary submovements are also considered the model by Meyer (Equation 11) is considered most appropriate. In applied sciences such as ergonomics and human computer interaction, the logarithmic model and particularly the one by MacKenzie (Equation 6) is the one that is used. This is due to the spatially constrained nature of the tasks. Logarithmic models are used for tasks where accuracy is an important factor such as when interacting with user interfaces. The models that focus on temporally constrained tasks are however useful in unravelling the nature of movements and the way they are performed. Verification of the existence of submovements and their nature is largely due to studies on temporarily constrained tasks. Figure 18 presents the competing models.

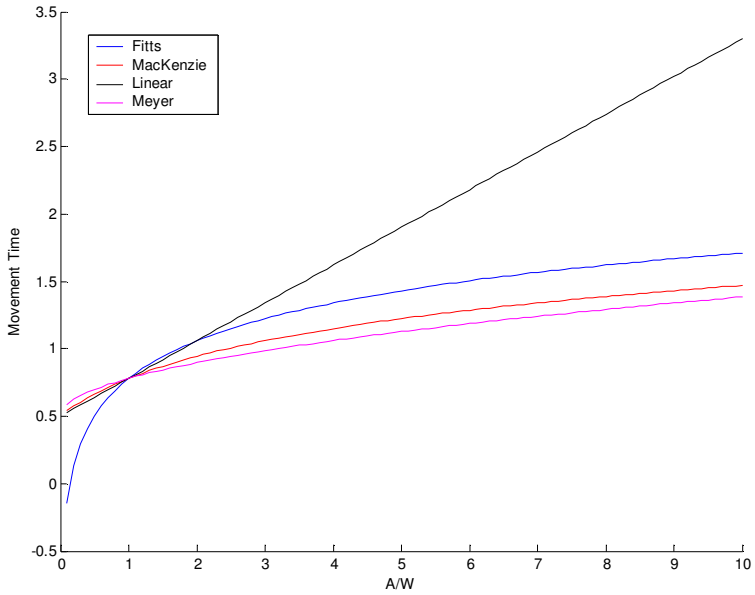


Figure 18. A graph showing movement time predictions for the competing models, for model values of intercept = 0.5 and slope = 0.28.

As can be observed in Figure 18, the models intersect at the A/W value of 1. For values lower than that the models predict movement times (from faster to slower) as Fitts', Linear, MacKenzie and Meyer. For A/W values higher than one the situation is the opposite and the models rate with respect to their movement predictions (from faster to slower) as Meyer, MacKenzie, Fitts' and Linear. The movement time predictions of the models are very similar for indices of difficulty up to approximately three. For lower values of intercept, the similarity extends to higher ID values. However, as the slope increases the differences of the models become more pronounced. Based on this fact, one can expect that it is easier to characterize a task based on an appropriate model when the experimental task can be performed at high A/W ratios or the slope values are high, implying relatively low index of performance values.

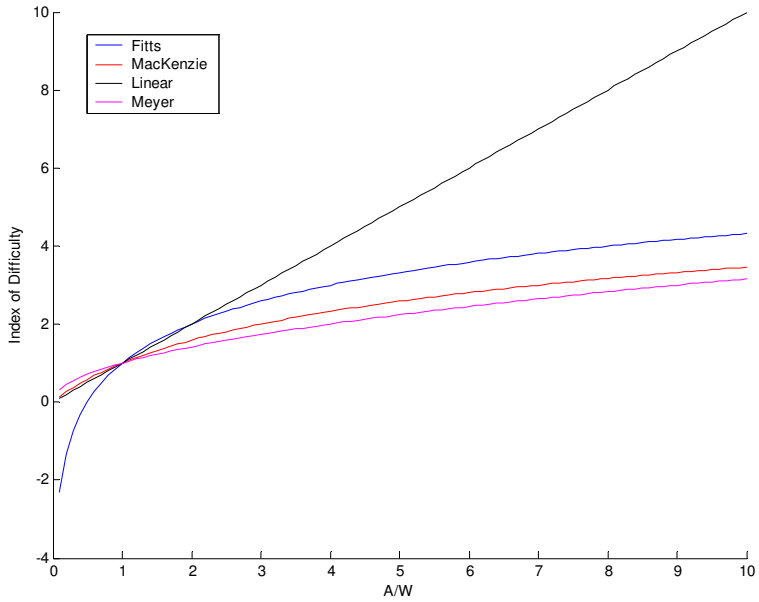


Figure 19. A graph showing Index of Difficulty predictions for the competing models, for model values of intercept = 0.5 and slope = 0.28.

Figure 19 presents the Index of Difficulty predictions of the models. The situation is analogous to the time predictions, with task predicted to be faster being rated at lower Index of Difficulty values. The figure also shows the negative index of difficulty prediction of Fitts' law for A/W values of less than 0.5.

Of considerable importance in the performance of target acquisition tasks is the role of practice. Citing all the studies on the effect of practice is out of the scope of the thesis. It is however stated, that researchers agree that movement time is reduced and accuracy of movement improved as a result of practice [95]. Both movement time and accuracy ratings improve after training and performance can be twice as good compared to that observed in the early stages of task performance [134]. It is also the case that the duration and span of the secondary submovements decreases while the duration and span of the primary submovements increases as user's become better trained in a certain movement task [34, 95, 96].

To explain the effects of practice it has been argued that a performer progresses from a closed loop mode of control that is dependent on the processing of online feedback to a more centrally driven mode of control that is less dependent on afferent information. However, recent research [34] indicates that this is not the case. Participants are moving in the same pattern and secondary submovements still occur no matter what the practice level is [82, 95]. Surprisingly enough, submovements also occur when visual feedback is deprived after task initiation [82, 95, 134]. It is quite important to note that both in temporally constrained tasks as well as Fitts' law in spatially constrained tasks Equation 11 applies under visual feedback deprivation as shown by Meyer as well as Wallace & Newel [82, 124]. Researchers used visual

feedback deprivation to identify whether visual feedback is important for the primary or the subsequent submovements or for both. Recent research agrees that with respect to the secondary submovements visual feedback is absolutely necessary [82, 95, 96, 134]. A large number of misses occurs when feedback is deprived when secondary submovements are executed. Although the variation in the primary submovement might not be great as the one of secondary submovements [82], recent research indicated that there is online processing of visual feedback during the first submovement as well [96, 134], however this is affecting the overall movement to a lesser extent.

The observation that corrective submovements still occur when visual feedback is deprived from participants is very interesting. The exact explanation of why this happens is however, largely unclear. It has been argued that corrective submovements [31] may be controlled either based on kinaesthetic feedback (mostly in physical movement studies) or due to a certain movement pre-programming that occurs prior to movement initiation. It has also been argued that the performance of secondary submovements in the absence of visual information is due to a time minimization pattern that participants follow in such a case, and therefore movements are not performed slowly and carefully, but rather through a hasty primary submovement that is followed by secondary submovements.

4.6 Conclusions

This chapter presented the methodology and the theoretical background that is currently used in evaluating aimed movement, a general category that includes pointing to visual targets. It is interesting to note that in human computer interaction, tasks are spatially constrained and therefore the logarithmic models are best suited for their modelling as found in a plethora of studies. The regression parameters (intercept and slope) that result from fitting the logarithmic models are used to uniquely assess the efficiency of a particular interaction technique. Based on the slope it is possible to calculate the Index of Performance, a parameter that indicates how a particular interaction technique behaves under different levels of task difficulty. The higher the Index of Performance the more insensitive is a certain task for increasing Indices of Difficulty. When reaction time and selection time are included in the time measurements, the intercept value provides information on reaction time and on the efficiency of the selection procedure associated with the task. When this is not the case, intercept values should be close to zero.

It is quite surprising that the effect of target width has not been considered explicitly in human computer interaction studies. In psychology research it has been found to affect both time and the accuracy of movement [82, 96], however even in this context no detailed investigation has been performed. This is therefore an interesting issue that is extensively investigated in the thesis.

Finally, of great importance are the observations on the fine nature of aimed movement. Experiments in psychology identified that aimed movement consist of a primary followed by secondary submovements. The primary submovement brings participants close to the target and the secondary

submovements fine tune the movement with respect to the target region. The rather insensitive nature of the primary submovement to detailed feedback (as appears from studies on visual feedback deprivation) makes deictic spatial audio target acquisition a promising venture. However, the importance of detailed feedback on the accuracy of the secondary submovement might compromise the success of the task given that auditory localization is blurred and less accurate than visual localization. As was explained in Chapter 2, the perception of sound direction in audio environments is poor compared to the visual space perception. In particular, evidence was provided that localization error and the lack of support for secondary submovements due to the absence of a perceivable border separating the target from the background are inherent in spatial sound perception. It is therefore evident that to result in spatial audio target acquisition of comparable speed and accuracy to vision, feedback has to be provided.

In short, this chapter provides the following points that are relevant to the thesis, in particular with respect to Research Questions 1 & 3:

- The success of an acquisition task depends on the adequate support of a primary (ballistic) movement followed by corrective submovements
- The wide applicability of Fitts' law implies that it might provide an appropriate model for a spatial audio target acquisition task (with adequate support for primary and secondary submovements)
- An acquisition task where the target is constrained horizontally but not vertically can be modelled as a one dimensional acquisition task
- A spatially constrained one dimensional task as the one examined in the thesis (see Chapter 7) is likely to conform to a logarithmic model

5 An initial investigation into spatial audio target acquisition

5.1 Introduction

This chapter presents an initial investigation into pointing interaction with a spatial audio display. The investigation targets mainly Research Questions 1 & 2. It investigates the feasibility of the spatial audio target acquisition task and effect of feedback. It also aims to observe how interaction accuracy is influenced by different pointing gestures and obtains an initial estimator on display target size as a function of the pointing gesture used.

5.2 Rationale

To construct the first experiment, it is necessary to provide:

- A spatial audio display design
- Selection procedures
- A method to estimate the accuracy of selections

The display used in the experiment will place sounds on a plane around the user's head at the height of the ears to avoid problems related to elevation perception, as was explained in Chapter 2. The whole circular area around a user's head is used. Spatial audio rendering is done using non-individualized HRTFs due to the availability and ease of use of this technique. As mentioned in Chapter 2, using individualized HRTFs requires measuring each individual's HRTFs. Such an option requires specialized equipment and space and cannot be thought to be a readily available option for the mass population at least at the time of the development of the thesis.

A key issue in 3D audio design is the number of sources that can be presented simultaneously. It has been shown that performance in identification, monitoring and intelligibility tasks degrades as the number of audio display elements increases [81] when sounds stem from the same point in the display. Spatial separation, however, forms a basic dimension in auditory stream segregation and thus can possibly increase the number of sources users can deal with. In the experiment, just one sound source is used, as an estimation of selection angles is desired for the simplest case, before more sophisticated sound designs are considered later in the thesis.

The experiment features two conditions, one where the experimental target area is marked with feedback and a second where the area is not marked with feedback. There are three reasons for using feedback. The first is to provide additional cues for overcoming front back confusions. If the user is asked

to reach the feedback area and make sure he/her can hear the feedback sound, then it is reasonable to expect that the front back confusion phenomenon will be minimized, even when misperceptions occur. The second is that feedback will explicitly denote target size and background target separation and in this way it could force participants select in smaller target areas. The third reason is that feedback is expected to reduce final positioning times and assist in the course of the experiment. In studies that employed active navigation in a sound field with no feedback, as for example in Loomis [68], users could home to the target relatively fast but increased final positioning times were observed. Finally, effective target widths are estimated at different angles around a user's body. This is done to identify possible selection effectiveness shortcomings due to motor deficiencies.

Because of the fact that the display uses sounds all around the user's head and also because non-individualized HRTFs are used, an exocentric design is chosen for this experiment and the effectiveness of target acquisition is examined with and without feedback. An exocentric design is interesting from an HCI point of view because it offers the possibility of sonifying objects of interest to the user that lie in a fixed position in the world, thus assisting navigation tasks as explained in Section 3.3. In addition, in an exocentric design, if users are asked to select a sound by physically turning to it, target sounds will be lying in the area of maximum localization accuracy at selection thus minimizing the effect of localization error. In this way problems related to the variation of localization accuracy around the user's head could be overcome. Finally, an exocentric design due to the fact that it supports active listening provides means of disambiguating confusions. According to this rationale, the selection procedure that is examined requires the user to turn so as the sound source appears to be coming in front of him/her and then he/her is asked to perform a selection gesture. Based on the literature, it is expected that users will be able to disambiguate front back confusions and select sounds with minimum effect of localisation error. Three different soundscape browsing and sound selection gestures are investigated in the experiment. The first is browsing using the user's head and selection by nodding, the second browsing by physical movement of the user's hand and selection by 'clicking', and the third browsing by moving a stylus, that controls a virtual pointer on a touch tablet and selecting by clicking a stylus button. More on the selection procedures are provided in Section 5.4. These gestures have been used in different contexts in Spatial Audio Display research as was mentioned in Section 3.5.

Finally, an adaptive psychophysical method was used to estimate the angle interval that results in 70% of a user's selections being on target. The reason for this is that such procedures have been successfully used for a long time for the estimation of appropriate values of physical quantities that result in certain subjective impressions. Such procedures are useful in answering questions such as: What is the appropriate value of a physical quantity that would result in a certain subjective impression with a certain probability? Such a question in the context of this experiment is transformed into: What is an appropriate value of target size so that people will be able to select on target with a certain probability? More

specifically a two-down one-up method is used. Section 5.3 provides the necessary background on adaptive psychophysical methods.

A study in a similar context is the one of Walker [123], where it was found that the area that was associated to an acoustic beacon by the system affected auditory navigation performance. However, this study was concerned with navigation in virtual auditory environments and did not involve physical pointing to a spatial audio source.

5.3 Adaptive Psychophysical Methods

A number of methods have been developed in psychophysical research in order to estimate stimuli values that satisfy a variety of perceptual properties with a certain probability. The goal is to support systematic exploration within sensory systems of the limits of detection and discrimination among similar and confusable stimuli. In addition it is usually desired to obtain general measures of behaviour that are subsequently interpreted as an indirect measure of perception. The most suitable methods for such measurements are adaptive methods. Adaptive methods have the advantage that they converge at the stimuli value of interest, thus avoiding the necessity to evaluate performance at fully sampled stimuli values.

Adaptive methods are characterized by the fact that a stimulus is adjusted depending on the course of an experiment until a value that satisfies some desired properties with a certain probability is reached. They result in measures of performance on psychophysical tasks as a function of stimulus strength or other characteristics. The result constitutes what is called a psychometric function [66]. The psychometric function provides fundamental data for psychophysics, with abscissa being the stimulus magnitude and the ordinate measuring the subjective response. In our days, the most commonly used methods are the Parameter Estimation by Sequential Testing (PEST) procedure [117, 118], the Maximum-Likelihood Adaptive Procedures [66] and the Staircase or Up-Down procedures [67].

In PEST procedures, stimuli are presented for a certain number of times at a fixed level and then a statistical test is performed to check whether performance at that level is better or poorer than the targeted performance level. If this is not the case stimuli level is adjusted, based on a predefined step size and the procedure is repeated. Stimuli is increased if performance falls below the targeted level and decreased if performance is above the desired level. Step size is also adjusted (reduced) during the procedure in order to obtain more reliable estimates. Modern PEST procedures use data from all trials to construct a psychometric function that is subsequently used to estimate the optimal step size for the next trial.

In Maximum Likelihood Methods, the performance estimates that are obtained by standard PEST procedures are supplied to an algorithm that attempts to fit a psychometric function to them. The new stimulus value is then determined not based on step size but rather based on the prediction obtained by the psychometric function. New data are used to update the psychometric function at each experimental trial. A maximum likelihood algorithm is typically used for fitting the psychometric function.

Finally, a widely used psychophysical method is the Staircase or Up-Down method. Up - Down procedures work by setting the stimulus to a certain level at the beginning of an experiment and then decreasing or increasing the stimulus based on the observation of a specific pattern in the subject's response. The phenomenon that occurs when the direction of stimulus change is reversed is called a reversal. Up-Down methods that decrease the stimulus after a valid answer and increase stimulus after an invalid answer converge to the 50% point of the associated psychometric function. A point of this function that corresponds to 50% would imply that at this stimulus level, 50% of the answers would be expected to be 'valid'. The drawback of this procedure is that it is strongly depended on a participant maintaining a stable response criterion. Fluctuations of the response criterion will lead to large fluctuations in the threshold estimates. To overcome this problem, psychophysical methods often aim at estimating as threshold the point at which the probability of correct responses is halfway between perfect and chance response, namely the 75% point. By altering the rule of stimulus change, different points of the psychometric function can be estimated. This manipulation of the rule constitutes what has been called transformed up-down methods as discussed by Levit [67]. The point of the psychometric function that the adaptive procedure will converge can still be calculated based on the fact that the probability of a 'valid' or an 'invalid' sequence will equal 0.5. For example, if the stimuli decrement is triggered by two correct responses, then the method is expected to converge at the 70.7% point of the psychometric function. By adjusting the rule, other points of the psychometric can be estimated. However, full sampling of the function is often hard to obtain due to the large number of experimental trials required. Kaernbach [60] proposed a method to obtain estimates other a target performance level, that is based on a simple rather than a transformed up-down method. Rather than adjusting the pattern for required for a stimulus adjustment, Kaernbach proposed using the simple Up-Down pattern but using a different step size for stimulus reductions than the one used for stimulus increments. In such a case, the targeted performance level p, can be used to estimate the ratio of downward to upward step size change r, using the formula $r = (1-p)/p$. The method is more efficient compared to complex transformed up-down procedures, however, the benefit is rather small for simple ones (as the two down – one up) and in addition there is the drawback that the large difference in step sizes might result in participants being able to anticipate stimulus changes and adjust their responses accordingly.

In the experiment a two down one up procedure is used to estimate angle intervals that will allow efficient selection of a spatial audio source. This is done due to the fact that such a staircase method is best suited for providing initial information on a psychometric function that is otherwise undetermined. Indeed, as mentioned by Leek [66], the maximum likelihood procedures involve fairly complex stimulus placement rules and in cases development of threshold estimates from the tracking data. In addition they require the assumption of a particular form of the underlying psychometric function, which is not well established for some psychometric tasks. The PEST method in addition, is expected to converge to the

appropriate stimulus value, in a relatively large amount of trials, that is in general more than the one required by a simple Up Down method.

The method is expected to converge at an angle interval (target width) where 70.7% of the selections will be on target. Although the estimated point of the psychometric function is expected to provide optimal performance, the two down one up procedure is preferred from other methods for reasons of efficiency. As will be seen in the experiment description, a number of up down procedures will be required to decide on minimum target widths that will allow efficient selections at different positions. For this reason, minimizing the time requirements of the experiment is of particular interest in this study.

5.4 Experiment

An experiment was designed to answer questions on the feasibility of spatial audio target acquisition and the role of feedback as well as the following ones:

- What is the minimum display area needed for the effective selection of a sound source?
- How is this area affected by the selection gesture, and what selection gesture is the most accurate?
- Which gesture was subjectively the easiest and most comfortable to perform?

The experimental task involved two stages. In the beginning, users had to orient the sound to their front using a browsing gesture. Subsequently, they had to perform a selection gesture associated with each browsing gesture to select the target. There were two participant groups, one where the target area was feedback marked and another where no feedback was given when the users were on the target area.

The three browsing gestures were: browsing with the head, browsing with the hand and browsing using a touch tablet. These gestures differ with respect to how common they are in everyday life. The first is the normal way humans perform active listening, with the position of the sound being updated as the user's head moves, so should be very easy to perform. The second is more like holding a microphone and moving it around in space to listen for sounds. The location of the sounds in the display is updated based on the direction of the right index finger. Direction is inferred by a 2D vector defined by the position of the head and the position of the index finger of the user. The third gesture can be thought as an extreme, in the sense that it cannot be mapped to a real world case. The user moves a stylus around the circumference of a circle on a tablet (the centre of the tablet marks the centre of the audio space) and the position of the sound source is determined by the stylus direction with respect to the centre of the tablet. In early pilot testing this type of sound positioning proved to be confusing if a user was to start a selection from the lower hemisphere. This was due to the fact that sounds moved as if the participant was looking backwards, although the participant was actually looking forwards. For this reason, left and right were reversed in case the user began browsing in the lower hemisphere. By doing this, the optimal path to the next sound could be found by always moving on the circle towards the direction in which the sound cue was perceived to be stronger.

The selection gestures were: nodding with the head, moving the index finger as if clicking a non-existent mouse button, and clicking a button available on the side of the stylus to indicate selection. In this experiment, three combinations of the above were examined: browsing with the head and selecting by nodding, browsing with the hand and selecting by gesturing with the index finger, and browsing with the pen on the tablet and selecting by clicking. Head/hand/stylus tracking was used to update the soundscape in real time to improve localization performance.

5.5 Stimuli and Apparatus

The aim of the experiment was to look at how the minimum angle interval that allows efficient selection of an audio source varies with respect to direction of sound event and interaction technique used. A single target sound was used that was placed in one of eight locations around the participants' head (every 45° starting from 0° in front of the user's nose) at a distance of two meters. This stimulus was a 0.9 second broadband electronic synthesizer sound, repeated every 1.2 seconds.

For the participant group that received feedback, very simple audio feedback was used to indicate that the user was within the target region and would select the sound source successfully. This was a short percussive sound that was played repeatedly while the user was 'on target' (i.e. within the current selection region) to assist each user in localizing the sound. This was played from the direction of the target sound. Sounds were played via headphones and spatially positioned in real time using the HRTF filtering implementation from Microsoft's DirectX 9 API. Sound positions were updated every 50msec.

To perform gesture recognition and finger tracking a Polhemus Fastrack was used that provides position and orientation data, and two sensors (see Figure 20). One sensor was mounted on top of the headphones to determine head orientation and help to recognize the nod gestures. A second sensor was mounted on top of the index finger to determine the orientation of the hand relative to the head and to recognize the clicking gesture in the hand condition. A Wacom tablet was used for the tablet condition. Nodding and clicking was calculated using velocity estimators inferred from the position data.

Figure 20. Selecting a virtual 3D audio source in the hand pointing condition.

5.6 Experimental Design & Hypotheses

The experiment design involved a between subjects factor (whether participants received or did not receive on-target feedback) and two within-subjects factors with each participant using each of the three interaction techniques in a counterbalanced order. There were two within-subjects independent variables: sound location (eight different levels) and interaction technique (three levels). The dependent variables were absolute deviation from target in degrees and effective selection angle also in degrees. Participants were also asked to rate the three interaction techniques used for browsing and selecting on a scale from one to ten with respect to how comfortable and how easy to use they found them.

5.6.1 Hypotheses

Before the experiment it was hypothesized that

1 the effective selection angle, the ease of use and the comfort that would be experienced by each subject would be affected by the gesture used.

2 the effective selection angle would not vary as a function of the position of the target sound due to the fact that selection would be performed when the target sound was in front of the user and thus localization error would be the same irrespective of target sound position.

5.7 Experimental Task

The participant's task was to browse the soundscape using the indicated browsing gesture until the sound was in front and then select the target sound using the associated selection gesture. The target sound repeated until the participant performed a selection. Upon selection, the stimulus was presented in a

different location randomly out of the set of available positions. The whole process was repeated until all up-down methods for each position converged. According to the up-down rule the target width was varied between trials; it was reduced after two on-target selections and increased after one off target selection. The step was initially 2° but was halved to 1° after the third reversal occurred. It should be noted that participants were unaware of this process; they were instructed to perform selections based only on audio feedback and localization cues. Participants stood wearing the headphones and tracker. They could turn around and move/point as they wished and were given a rest after each condition. The experiment could not be conducted in a fully mobile way with users walking due to the tracking technology needed for gesture recognition – participants had to stay within range of the Polhemus receiver.

5.8 Procedure & Participants

Participants were trained for a short period before being tested in each condition to ensure they were familiar with the interaction techniques. They performed eight selections using each interaction technique before embarking on the experiment. Prior to testing, participants' localisation skills were checked to rule out hearing problems and to familiarise them with the sound signal they would hear. During this 3D sound training, participants were asked to indicate verbally the direction they had perceived the sound source was coming from. The experimenter subsequently corrected them in case they were wrong and tried to direct their attention to the relevant cues. Thereafter, they embarked on the experiment and were tested in all three interaction techniques according to the order that was dictated by the experimental design. The experiment lasted approximately one hour. The plan was to test twenty four participants, twelve with and twelve without feedback. Only four participants were tested in the non-feedback case since the task proved to be too difficult to perform. Twelve participants took part in the feedback case: five females and seven males with ages ranging from 19 to 30 years. Participants were asked whether they were facing any hearing deficiencies and if they did they were excluded from the experiment.

5.9 Results

The analysis presented here is based on the raw data that are available in the associated section of Appendix 1. The up-down method was expected to converge on the point of the associated psychometric function where 70.7% of the selections would be on target. To estimate this point the angle intervals as these were updated by the up-down rule were averaged. Averaging included only the angle intervals that occurred after the second reversal. The up-down method converged only for the feedback case. Responses in the non-feedback case were such that the up-down method did not converge and thus no results are presented for this case.

A 3x8 two factor ANOVA was performed to examine whether sound location and interaction technique affected the effective selection angle. Sound location was not found to have a significant main

effect (F(7, 77) = 2.241, p = 0.121). However, there was a significant main effect for interaction technique (F(2, 22) = 10.777, p < 0.001). There was no interaction between location and technique. Pair-wise comparisons using Bonferroni confidence interval adjustments showed that the tablet condition was significantly more accurate than the other two techniques, but no significant differences were found between the hand and head. Figure 21 shows the mean effective angle intervals for the three interaction techniques with respect to direction of the sound. Post-hoc one way ANOVAs showed that sound position did not have a significant main effect in the case of the tablet and the head but did have one in the case of hand pointing, F(7,77) = 56.127, p = 0.036.

These results define the one side interval around a source. To give an example of how these data could be applied, if an exocentric 3D audio user interface (enabled with active listening) using audio feedback and controlled by a stylus on a touch tablet, was developed, the designer should allow at least 4° on *each side* of a sound positioned at 90° relative to the front of the user so that a user would be able to select the sound with selection success rates in the order of 70.7%.

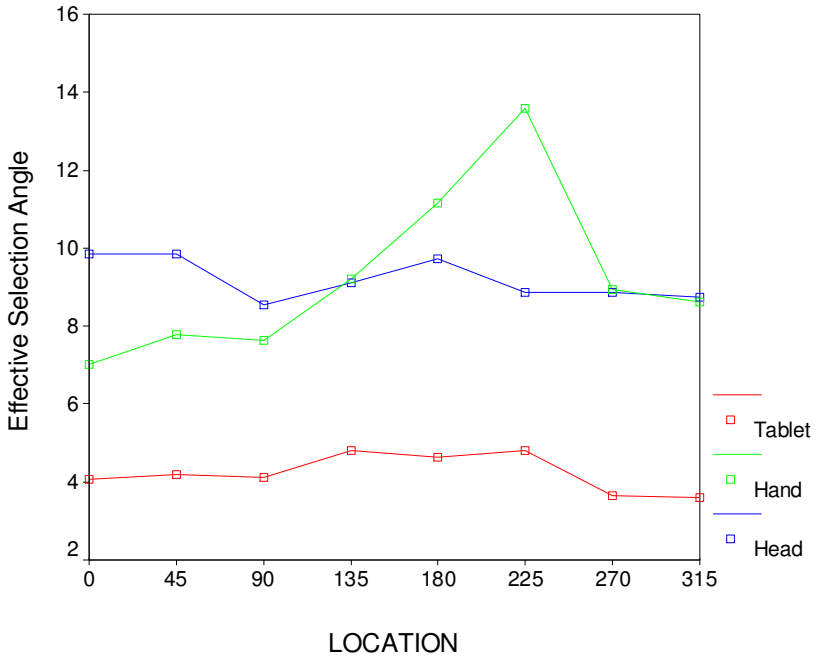


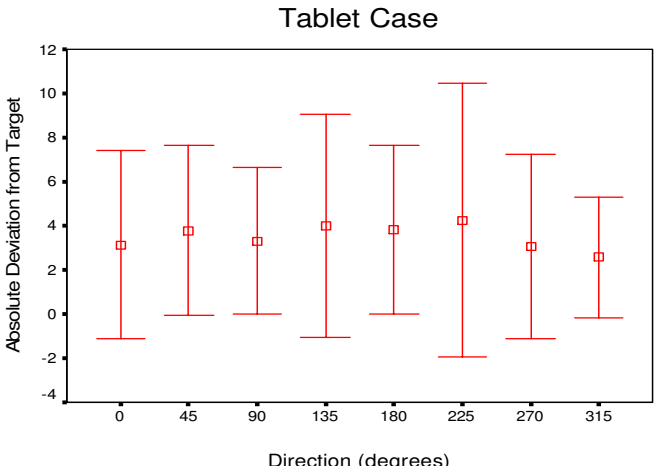
Figure 21. Effective selection angle as a function of sound direction for each interaction technique.

In addition, to the data provided in Figure 21, means and standard deviation values are provided in Table 3, to complete the picture.

	Tablet		Hand		Head	
	Mean	Std	Mean	Std	Mean	Std
0°	4.06°	2.46°	7.01°	3.03°	9.85°	5.68°
45°	4.19°	2.94°	7.78°	3.68°	9.85°	5.55°
90°	4.11°	2.52°	7.62°	5.02°	8.53°	3.86°
135°	4.80°	2.71°	9.20°	6.60°	9.10°	3.69°
180°	4.62°	2.10°	11.15°	8.59°	9.71°	5.10°
225°	4.81°	2.46°	13.58°	8.53°	8.86°	4.4°
270°	3.65°	1.95°	8.93°	3.62°	8.86°	4.22°
315°	3.60°	1.96°	8.61°	4.50°	8.74°	4.41°

Table 3. Mean Effective Angle and Standard Deviations for all interaction techniques.

The absolute deviations from target of the users' selections were also analyzed. Ninety measurements for all different directions were used. A 3x8 two factor ANOVA showed a significant main effect for interaction technique ($F(2,192) = 31.107$, $p = 0.001$). Direction also had a significant main effect ($F(7,672) = 4.025$, $p = 0.001$). There was a significant interaction between technique and direction ($F(14,1344) = 4.336$, $p = 0.001$). Pair-wise comparisons Bonferroni confidence interval adjustments showed that the tablet condition was significantly better than the others, but there was no significant difference between head and hand. With respect to the direction of the sound event, direction 225° was significantly different from directions 0°, 45°, 90°, 135°, 270°, 315° and direction 180° was different from 45°, 270°, 315°. Figure 22 illustrates mean absolute deviation from target and its standard deviation.



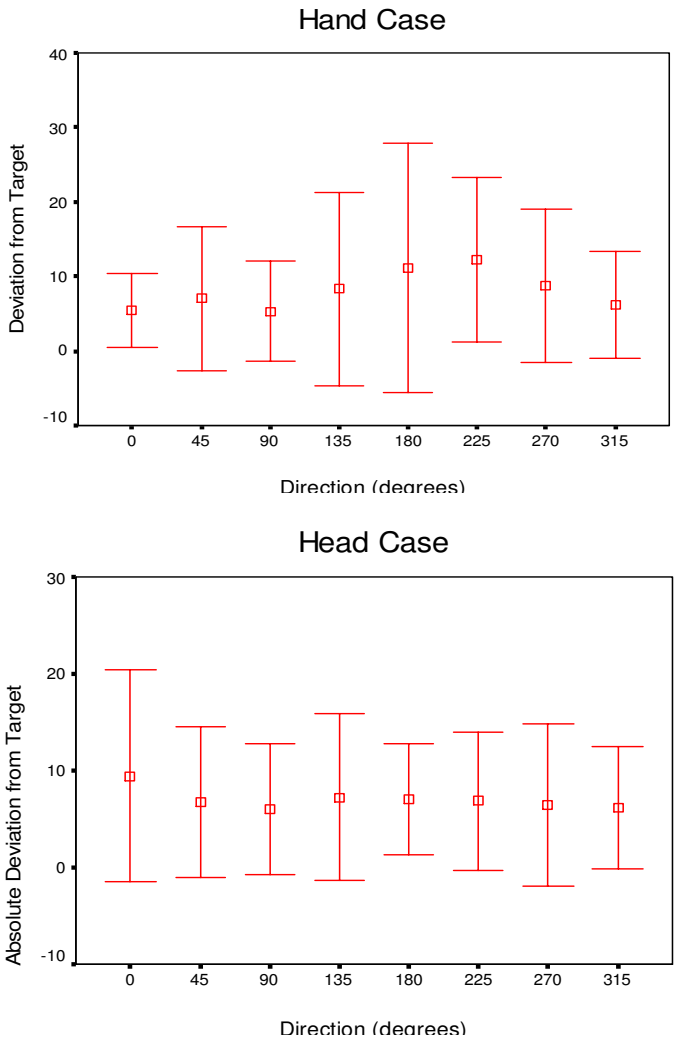


Figure 22.Mean absolute deviation from target and its standard deviation versus sound direction and interaction technique.

As mentioned, each participant was asked to rate each of the interaction methods in terms of how easy and how comfortable he/she found them to be, on a scale from 1 to 10. Figure 23 shows the means of the results for ease of use.

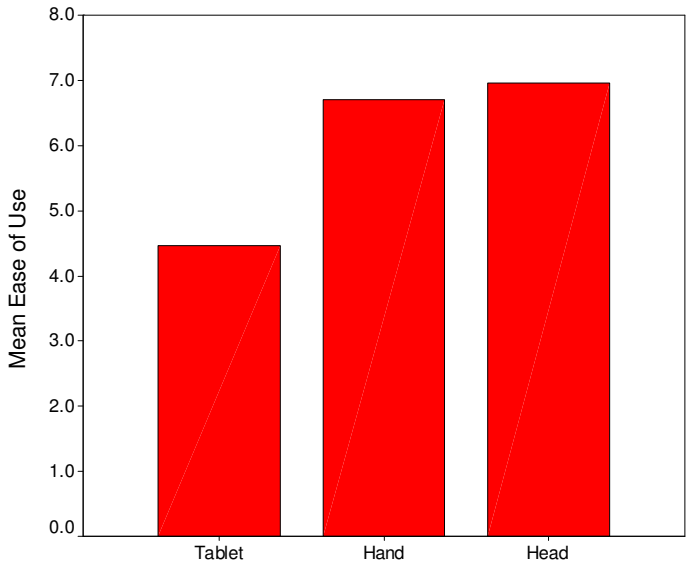


Figure 23. Mean ease of use ratings for each interaction technique.

A statistical analysis of variance showed interaction method to be a significant factor ($F(2, 22) = 5.8$, $p = 0.009$). Bonferroni t-tests verified tablet to be significantly harder to use, but showed no statistical difference between hand and head.

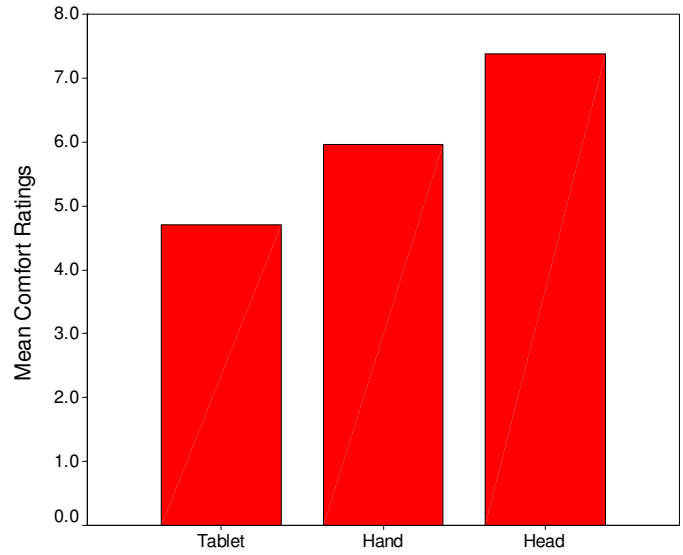


Figure 24. Mean Comfort ratings for each interaction technique

A similar analysis on how comfortable the use of the three devices was showed no significant difference between devices. Figure 24 shows comfort means for the three interaction methods. It should be noted that participants have performed a large number of selection using the three interaction methods

to allow the up-down methods to converge in the three different conditions and the eight different sound positions. In that sense, when observing the graphs, absolute values should be taken into account carefully. However, the ratings of the three devices relative to each other can be used to infer how they are ordered relative to each other with respect to ease of use and comfort.

5.10 Discussion

The results of the study showed that gesture interaction with a 3D audio source can be performed effectively in the presence of localization feedback. In the non-feedback case, with few exceptions the up-down procedure did not converge. Based on this finding, testing was not completed in this case. The fact that the procedure did not converge indicates that participants were not able to locate the target with confidence in this case. Such a finding can be attributed mainly to front-back confusions and inabilities to maintain a stable criterion on the perceived sound direction in the given experimental scenario. Indeed, given that participants did not receive feedback, target sound position was located all around participants and non-individualized HRTFs were used, high confusion rates and localization error would be expected with detrimental effect on the convergence of the Up-Down procedure. Indeed, the up-down procedure either diverged or unusable estimates of target size were obtained. For the small number of participants that were able to complete the procedure, angle estimates were similar to the ones obtained in the feedback case, however, experiment time was much higher. It should also be noted that the participants of the experiment had no experience in localizing virtual 3D audio sources.

In the presence of localization feedback, however, the novel browsing and selection methods that were tested were found to be effective. Users were able to perform active listening using the tablet, the head and the hand without any particular difficulty. In terms of selection effectiveness, the browsing and selection methods can be rated as tablet, hand and head. The ordering can be explained in terms of the resolution that the three different mechanisms provide. A stylus controlled touch tablet provides a much better minimum possible displacement compared to the head or the hand of a person. By constructing histograms of deviation data it was verified that the results of the up-down procedure would indeed allow 70% percent of the selections to be on target. It should be mentioned, however, that more reversals would result in having more accurate results. This was not possible to do since a within-subjects design and a short experiment duration was targeted by the experimental design. Experiment time was kept in the order of one hour to avoid effects caused by fatigue. Effective selection angles are likely to reduce with practice and improved feedback design.

When considering the three interaction methods, it would not be expected that the direction of sound would be a significant factor in the results of this study. This is due to the fact that users selected a sound when it was in front of them. Such a prediction was verified in the effective angle case where location was not found to have a significant main effect. However, in the deviation analysis, certain angles were different from others. This was mostly around the direction of 225° degrees. The reason for

this difference can be described by the mechanics of the browsing and selection modalities. A closer look at the graphs and Table 3 reveals the technique that caused this difference was browsing by hand. As was observed during testing, some right-handed participants found it difficult to point to that location, if they had not turned their bodies first (they had to reach around their body causing them to stretch, reducing the accuracy of their selections). A significant number of participants indeed tried to point without turning their bodies, a result that influenced the accuracy of the browsing and selection processes. The variation in the selection strategy can also be verified by the larger standard deviations that occurred for this particular region.

By analyzing how the ease of use scores are ordered, it is found that users find browsing the orientation updated sound space to be equally easy either using the head or using the hand. The touch tablet however, although more accurate, was not rated highly. This can be associated with the unnaturalness of the browsing process. In the other two cases, participants used a natural process for browsing the space, such as moving their heads or simulated one by moving their hand in a synchronous way with their head. The unnaturalness of the process however, is tightly coupled with the exocentric nature of the display. In an egocentric display where sound position remains stable relative to the listener, such a result might not be verified.

When considering the effective angles, it can be observed that if accuracy was the only factor to be taken into account, an audio user interface could be constructed having all eight sounds locations, and possibly more.

5.11 Conclusions, Guidelines & Future Directions

Experimental hypothesis 1 it was found to be valid, gesture affected the effective selection angle, the comfort and the ease of use participants experienced. Experimental hypothesis 2 was found to be partly valid with exception of selection using the hand gesture. This is consistent with the findings of Oldfield and Parker [89, 90] where pointing to sounds in the back was found to result in increased motor error. In our case significantly higher effective selection angles were also observed when the target sound was on the back of the participants.

The experiment showed that gesture interaction with an exocentric spatial audio display is a feasible task in the presence of on-target feedback. In such a case, it was observed that participants were able to orient themselves with respect to the sound source and select within reasonable limits of accuracy. Without feedback however, problems related to front-back confusions and the relative inability of listeners to orient themselves successfully within the target region resulted in ineffective interaction. Three different gestures for browsing and selecting in a 3D soundscape were examined and their effectiveness in terms of accuracy was assessed. Browsing and selecting using a touch tablet proved to be more accurate than using a hand or a head gesture. However, browsing and selecting using the hand or the head were found to be easier and more comfortable by the users. Effective selection angles that would

allow efficient selection were estimated for each interaction technique and at eight sound locations around the user using an adaptive psychophysical method. The results showed that these different interaction techniques were effective and could be used in a future mobile device to provide a flexible, eyes free way to interact with a system. It was found however, that orientation cues from an actively orientated display were not sufficient and in addition that conventional techniques using virtual pointers produced high cognitive loads within such a context. With the goal of designing a spatial audio selection task, it is therefore necessary to broaden the area of research to include egocentric displays and other localization cue options. Furthermore, restricting the display space to lie only in front of the user might also help in alleviating confusions in a more effective way. In addition, an evaluation method has to be devised to include efficiency measures and more formal ways to assess user satisfaction. Indeed, in the experiment, large time differences were observed between the feedback and non-feedback case. These were not however, quantified in the experiment as this served as an initial study into the feasibility of the interaction techniques under examination. Given however, the verification of the effectiveness of the interaction techniques under examination and these initial observations, the next experiments will have to explicitly evaluate the efficiency and user satisfaction aspects of gesture interaction with a spatial audio display.

Finally, as future research direction for developers of exocentric interfaces where selection is important, hybrid methods of browsing and selecting could be examined. For example, browsing the virtual soundscape using the head but selecting using the hand, could result in a solution that is optimal from the point of user satisfaction but also efficient from a scalability point of view. The same could be true for browsing using the head and selecting using a touch tablet for example. However, these assumptions have to be tested and verified experimentally to ensure their usability and reveal possible shortcomings.

With respect to Research Question 1 this experiment showed that if feedback marked areas are used in a spatial audio display the task of selecting a spatial audio target becomes feasible. With respect to Research Question 2, the study showed that the gesture used significantly affects the accuracy of user selections. Based on the experimental results a number of guidelines for developers are provided:

- When it is desirable to provide front/back cues in an exocentric display to listeners using non individualized HRTF functions, it is necessary to mark the target area with feedback so that the selection task becomes usable.
- Natural options of browsing an exocentric display, such as by head movements, are easier and more comfortable to perform by users.
- When selecting a feedback marked spatial audio element using a stylus operated touch tablet 9° around the target should result on 70.7% selection success rate. This number becomes 16° and 18.5° for selecting using the hand and head of the user respectively. In the hand case, this number refers to selections in front of the user.

6 An investigation into deictic interaction in egocentric and exocentric displays. The effect of feedback cues and distracter sounds.

6.1 Introduction

In this chapter the methodology that was used in the previous experiment is extended and used to evaluate a number of designs that aim to enhance pointing efficiency using a number of feedback cues. In this way, a preliminary attempt to answer Research Question 3 is made. A distinction between egocentric and exocentric displays is made and two feedback cues are used in each of the display types to examine their effect on performance. The experiment is expected to provide the necessary insight for answering Research Question 1 and lead to the design of a spatial audio target acquisition task. Furthermore it is expected to provide more material for answering Research Question 2 by identifying the effect of feedback types that differ in a fundamental way.

In addition, this experiment is investigating the effect of distracting sound sources on interaction. This is done with respect to Research Question 2 in order to identify whether distracter sounds affect spatial audio target acquisition. Justification for this investigation is provided by the fact that in the real-world, an audio display is expected to feature more than one display element. In this sense, it is interesting to question how interaction will be affected.

6.2 Rationale

Based on the discussion in Section 3.3.2, it can be concluded that egocentric and exocentric designs have some fundamental differences in terms of application areas and it is also expected that some notable differences would be expected in interaction patterns. Given that both display types are considered useful each in its own scope, it is necessary to see whether these differences would be observed experimentally and whether they are possible to quantify. Furthermore, both display types may suffer either from effectiveness or efficiency related deficiencies. The results of the experiment presented in Section 5, also indicated the important effect feedback has on interaction. It is therefore necessary to investigate whether feedback cues imposed on these basic designs can counter these problems. Feedback can however, be provided by a variety of methods and the merits and drawbacks of each technique have to be investigated. In particular, it is necessary to examine how performance is affected by a number of prominent feedback cues, such as loudness, orientation update and the simple timbre cue that was used in the previous experiment.

In order to gain more conclusive results, the effectiveness of feedback has to be evaluated against variable display element numbers. In this way, the effect of distracting sounds in the display on

interaction can be examined. Although similar effects have been widely investigated from an intelligibility point of view, the effects on interaction have not been evaluated in the past. Based on these observations the aim of this experiment is defined as to evaluate a number of prominent feedback cue types and study how interaction assisted by each feedback cue is affected by increasing the number of display elements in egocentric and exocentric interfaces.

The chapter provides background on the effects of simultaneous presentation of audio streams on interaction as well as on feedback design, before extending the evaluation technique and proceeding with the experiment.

6.3 An introduction to the issues associated with simultaneous presentation of audio streams

The most important effects of display content on interaction with a complex spatial audio display are expected to emerge as the result of intelligibility problems and possibly problems related to increased user workload due to inappropriate sound design. The former possibility has been investigated in the psychoacoustics literature. The latter remains however, largely, unexplored and more research is necessary in order to develop sound design guidelines.

When interacting with a spatial audio display a user is faced with a complex audio environment where multiple sounds might coexist (including sounds from the real audio environment surrounding the user). From this point of view, interaction with a spatial audio display is highly associated with divided and selective hearing attention tasks [109]. Divided attention tasks are those in which the user must follow more than one information stream at a time. Selective attention tasks are tasks where attention is focused on only one of multiple information streams. For example, listening simultaneously to two speakers in a teleconference scenario is a divided selection task, since the user is required to understand the 'meaning' conveyed by both of the display elements. On the other hand, the task of selecting a target audio element is a selective attention task, since the user has to focus on the target element with the rest of the display elements acting as distracters. Intelligibility problems in both cases mainly stem from the phenomenon of masking.

Masking is defined as the process or the amount by which the threshold of audibility for one sound is raised by the presence of another sound [85]. When audio display elements are presented simultaneously masking can lead to one or more of them being inaudible if there is significant spectral overlap between them together with marked level differences [85]. In general, masking in spatial audio display is less of a problem due to the fact that both target signals and maskers might possess individual spectral and temporal structure that has been proven to assist auditory stream segregation. In addition, in spatial audio displays sounds are presented from different spatial locations. In such cases, the masked threshold is lower compared to when sounds are presented from the same locations. This phenomenon is

called *binaural release from masking* and is one of the advantages that spatial audio displays have compared to non-spatial audio displays [85].

Binaural release from masking has been used to explain the ability of the human auditory system to focus in one of multiple audio streams that are presented simultaneously, known as the 'Cocktail Party Effect' [4]. The individual differences of sound signals between ears have been found to be helpful in reducing the threshold of audibility for sounds presented in the presence of maskers. For all these reasons, performance in selective attention tasks is acceptable in spatial audio displays as long as the levels between display elements do not have big differences [109]. It should be noted that divided attention tasks are more demanding and the benefit from spatial separation is less than in selective attention tasks. As reported in [109], at 0 dB target to masker ratio participants performed at a success rate of 95% in the selective attention task but the success rate in the divided attention task was only slightly more than 70%.

In addition to the aforementioned masking type, also known as energetic masking, there is also the case of informational masking. Informational masking stems from the observation that high levels of masker uncertainty result in higher masked thresholds [71]. Given this observation it is reasonable to assume that consistency in the timbre of the display elements is an aid to a user interacting with a spatial audio display. In addition, spatial separation also improves performance and reduces the effect of informational masking [38]. It has also been found that previous knowledge of the position of the a target is beneficial to intelligibility performance in selective and divided attention tasks so keeping display elements fixed relative to the user may prove beneficial. Therefore, in a familiar spatial audio display the amount of informational masking is expected to be minimal.

According to the review, spatial audio displays seem to favor both types of attention tasks in terms of intelligibility and masking avoidance. However, the effect display 'clutter' has on interaction is not clear as most of the studies focus on intelligibility rather than performance. From this point of view, it is interesting to examine how and whether user performance would be influenced by variable levels of display content (as would happen in any real system).

6.4 Feedback Cues

Feedback can be either constrained in the display area that is implicitly assigned to the sound in focus or be provided by continuously updating a display parameter, possibly coupled to user movement. A simple way to provide feedback is by playing an external sound source whenever the user is on target. This type of feedback cue is going to be referred to as a timbre cue in the rest of the chapter. However, such type of feedback can be altered so as pitch, rhythmic or spectral variations of the sound representing the display element are played whenever it has focus. As has been found in the previous chapter, this approach is effective and can successfully improve selection speed and accuracy. Such a choice is not unnatural and it is justified by the fact that feedback has to be provided anyway to inform on the current

display state, for example to show whether a certain display element is in focus or that it has been selected etc. With appropriate design such feedback can also be used for the additional purpose of assisting users in disambiguating display element position and overcoming speed and accuracy related deficiencies. More specifically this type of feedback might be used to either improve final positioning time in an exocentric display or compensate for accuracy deficiencies in an egocentric display. Final positioning time is the elapsed time from when a user enters the target area to the moment a selection is made. Even when targeting visual targets, on-target feedback has been found to produce marked differences in final positioning times as found by Akamatsu *et al.* [3]. In this study, auditory, tactile, colour and all three combined feedback types were compared as a means to indicate that the pointer was over the target. An analysis on final positioning times gave a ranking of tactile, combined, audio, colour and normal. The differences in mean times were not pronounced but were significant and based on this study it can be concluded that feedback can improve final positioning times. Final positioning is also a significant problem in audio displays [68]. In an experimental study Loomis *et al.* asked participants to locate a sound by physically moving to it. Sound position was updated in real time using distance and orientation cues depending on the user's relative position with respect to the target. The authors found that people could quickly get to the target sound source however, there was a significant delay until participants were convinced that they were actually on target.

Feedback can also be provided by adjusting display parameters to give hints on the position of display elements. This is done using information about user position, obtained through orientation or position tracking equipment. An example of such a technique would be updating the loudness of the display element at which the user is pointing. Such a cue can be designed by means of a function that relates the attenuation applied to each display element to the user's distance from the display element. Continuous or discrete attenuation levels can be used, the latter done through mapping of attenuation levels to different ranges of user distance to target. A loudness based cue can guide the user to a target display element location since the loudness of the particular element will increase as the user moves closer to the display element. In addition, such a cue effectively adjusts the target to masker ratios in the display and as such it is helpful in the context of enhancing divided as well as selective attention. This is also important in mobile settings as it can help overcome problems of masking by display or other real world sounds. At very high attenuation levels the loudness cue is effectively reducing display population, since elements far from the user's pointing direction will be rendered inaudible. This might become problematic since continuous contact with display elements is not preserved.

An exocentric display implicitly provides orientation cues to the listener due to the on-line updating of sound position relative to user position. If a user is advised that a sound will be in focus whenever it is in front of the user, a display browsing and selection strategy can be devised that uses the readily available orientation cues. Given the applicability of such options in display design, it is

interesting to evaluate different movement-coupled feedback cues and rate them according to the benefit they bring to interaction.

6.5 Evaluation Methodology

As was said in the introduction the evaluation methodology is extended in this chapter to include more aspects of the usability of pointing actions. Performance is evaluated in terms of efficiency and effectiveness by measuring time and accuracy scores respectively, in a spatial audio target acquisition task. User satisfaction is estimated by using the NASA TLX workload questionnaire [54]. Using the questionnaire participants rated their subjective experience in a number of factors that are considered to influence subjective workload during interaction. These scores can be subsequently used to calculate the perceived workload in a certain interaction context.

Finally, two additional standard measures from the literature on evaluating pointing actions are introduced: effective target width and throughput. Throughput is defined as:

$$Throughput = \frac{ID_e}{MT}$$ **Equation 28**

ID_e being the index of difficulty and MT the movement time. Index of difficulty is defined as:

$$ID_e = \log_2\left(\frac{D}{W_e} + 1\right)$$ **Equation 29**

W_e is the effective target width calculated based on the standard deviation of measurements and represents the distribution of selection coordinates computed over a sequence of trials. It is calculated as:

$$W_e = 4.133 \times SD_x$$ **Equation 30**

SD_x is the standard deviation in the selection coordinates measured along the axis of approach to the target. To apply the above formulation in our study D is defined to be the angular distance participants had to move to reach the target, measured in degrees. The particular measures have been proposed for evaluating visual target acquisition, however here an attempt is made to extend their application to spatial audio selection tasks. This is justified because spatial audio is providing a directional cue and therefore the spatial audio target acquisition procedure can be considered similar to visual target acquisition. As far as effective target width is concerned the application is not questionable due to the fact that Equation 28 does not involve any terms that can be thought to relate to modality. With respect to throughput, it might be possible to question the appropriateness of the formulation Equation 29 as a measure of difficulty of a

spatial audio target acquisition task. It is indeed an open question whether the formulation of Equation 29 can be applied to spatial audio selection tasks. However, the thesis proceeds with using the formulation and uses it uniformly in this study for all feedback cues given, because it provides useful insight into their effectiveness. Chapters 7 and 8 consider in detail whether such formulations are appropriate for spatial audio target acquisition tasks.

6.6 Experiment

An experiment was designed to evaluate interaction in the presence of the discussed feedback cues, input being accomplished by means of a physical pointing gesture. According to the rationale the feedback cues are evaluated in a display with varied display populations to obtain information on whether their effectiveness is affected. The term display populations refers to the number of display elements that are presented in the display.

Interaction with the display is accomplished in this study using simple physical gestures that are recognized by the system using motion trackers. Physical gestures are a suitable solution for interacting in mobile contexts due to the fact they can be performed without using stylus or similar devices.

6.7 Experiment Design & Hypotheses

The Independent Variables are orientation update (between-subjects) that differentiates between the egocentric and exocentric interface, feedback type and distracter population (both within-subjects). Dependent variables are time to complete a trial, angular deviation from target, throughput and effective target widths as well as subjective workload. Participants were split into two groups: one with orientation update enabled in the display and the other without. The feedback type variable was introduced to accommodate the loudness and timbre cue. A void level was used to provide the control condition of direct pointing in egocentric and exocentric interfaces. The combination of orientation update and feedback type resulted in six different feedback cue combinations. The control condition was provided by the combination of no orientation update and no feedback cue and essentially represents direct pointing. The loudness and the timbre cues were tested with and without the orientation update cue to examine what is the effect of cue combinations. Display population was also tested as a within subjects factor with all participants tested in all available levels of display content. The maximum number of sounds in the display was seven including the target sound and the minimum just one, the target sound. The design of the experiment is presented in detail in Table 4.

Orientation Update	Feedback Type	Number of Display Elements
Yes (Exocentric)	None	1,2,3,4,5,6,7
	Loudness	1,2,3,4,5,6,7
	Timbre	1,2,3,4,5,6,7

No (Egocentric)	None	1,2,3,4,5,6,7
	Loudness	1,2,3,4,5,6,7
	Timbre	1,2,3,4,5,6,7

Table 4. Experimental Design. The independent variables, orientation update, feedback type and number of display elements and their associated levels.

6.7.1 Experimental Hypotheses

Before the experiment it was hypothesized that:

1 Interaction in the egocentric display would be faster but less accurate that interaction in the exocentric display

2 The timbre cue was expected to result in usable interaction in both cases, however its success would be compromised by the fact that no direct contact with the target element would be available in the beginning of each trial. However, the repetitive nature of the task was expected to smooth this problem

3 The success of the loudness cue was expected to be compromised by the fact that it does not directly provide directional information. It is therefore expected to improve accuracy compared to the egocentric case, however its success in the exocentric case depends on the extend participants would utilize the combination of the orientation and loudness cues. Furthermore, the lack of continuous contact with the target in the beginning of each trial was available is also expected to influence interaction in a negative way.

4 Distracter sounds were expected to influence performance in a negative way.

6.8 Stimuli & Apparatus

Participants were tested in a spatial audio display that is presented in Figure 25. Display content was constrained to a maximum of seven audio elements. Elements were positioned on the arc of a circle of radius of 3m starting from -70° and up to 70° with an inter-element distance of 20° at the level of the user's nose. The selection of a 3m distance was due to the fact that this sets the sounds in the 'far' field where spectral cues are rather uniform and do not depend on sound position. Interaction was accomplished by means of an Xsens MT-9B orientation tracker (www.xsens.com), which participants held in their hand. Headphones were used to present the sounds.

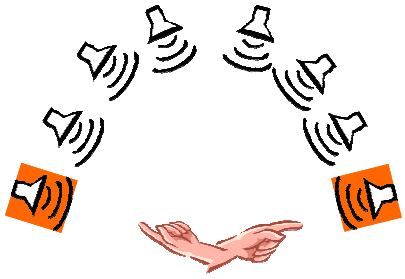


Figure 25. Experimental Task, Visualization of the hand of a participant alternating between the two targets while tested in the display.

When the display featured orientation update, sound positions were updated automatically based on the direction of the user's hand. The DieselPowerEngine (www.am3d.com) was used for spatial audio rendering of the display elements. The DieselPowerEngine was found to be better compared to the DirectX as the evaluation results presented in the Appendix showed. The loudness cue was implemented using a continuous bell-shaped attenuation function designed so that when a participant was pointing straight at the location of a display element neighbouring elements were played at half their original level. The shape of the function was such that sounds other than the focused and the neighbouring ones were played at zero level so that they were inaudible. Attenuation levels were continuous. To implement the timbre cue each display element was assigned an effective area of 20°. When using the timbre cue only the display element in whose effective area the user was pointing was audible. On entering the effective area of a display element, the associated sound was played from the beginning. This type of feedback is similar to the case where a different sound or a variation of an element sound is used to inform the user of a display element being in focus, with the notable difference that no continuous contact exists with the other target elements.

One target sound appeared in the display in each trial and this was a human voice saying 'Woohoo'. To provide a rather uniform sound material, animal sounds were used for the rest of the display elements that were used as distracters. The distracter population was chosen randomly for each particular trial out of a maximum of six sounds. These were the sounds of a kitten meowing, a puppy barking, a horse whining, a cockerel crowing, a cricket chirping and a hen clucking. The sounds were equalized to have roughly the same loudness. Sounds were HRTF-filtered in real time to provide the impression of them emitting from a certain position in space. Sounds were programmed to play as omni-directional sound sources and no directional characteristics were used. The sounds in the display repeated with a 400msec pause until a selection was made. Onsets and durations were not constrained.

6.9 Experimental Task

The experimental task was to select the target sound using the feedback cues that were available in each condition. To select the target sound participants had to point at its perceived direction and then rotate the hand slightly downwards to indicate selection. The target sound alternated between the leftmost and rightmost display slot in every other trial. This was done in order to minimize searching time and allow for the effects of the distracting sounds and the combinations of the feedback cues to appear. The number of display elements was selected randomly prior to a new trial out of a maximum of seven. Display element positions were filled from the direction of the target according to the number of display elements for each trial. A visualization of the experimental task was provided in Figure 25. The task is designed in accordance to the original Fitts' pointing task paradigm with the difference that distracter sounds were also played in between the target display elements and that stimuli were placed in an arc rather than in a straight line.

When the display featured the orientation update cue participants were asked to use the updated orientation cues and select a sound when it was in front of them. When using the loudness cue participants were asked to select when the sound was at its loudest. When using the timbre cue participants were asked to select when they could hear the target sound. When using combination of cues participants were asked to use both cues on their own initiative. A participants performing the task is shown in Figure 26.

Figure 26. Performing the task in Experiment 2. A user is illustrated selecting a 3D audio source.

6.10 Procedure & Participants

There was a counterbalanced testing order for the within subjects factors. Participants were trained briefly prior to embarking on each of the within-subjects conditions. This was done mostly to familiarize themselves with the cues available in each condition as well as with the experimental equipment. Pointing angle at selection and time to complete trials were recorded throughout the experiment. After completing each of the trial sets participants were asked to complete NASA TLX forms in order to estimate the subjective workload associated with each trial session. Twelve participants were tested 11 male and 1 female with mean age of 22 years. Participants had no previous experience with interacting with an audio display. They were paid 5£ for their participation. Participants were asked whether they were facing any hearing deficiencies and if they did they were excluded from the experiment. Total experiment time varied between 30 minutes and one hour.

6.11 Results

The analysis presented here is based on the raw data that are available in the associated section of Appendix 1. The results section is organized in three subsections. The first is concerned with the analysis of movement times, the second with the analysis of deviation from target scores and additional observations concerning throughput and effective widths as these were calculated for each feedback cue and their combinations. The third is concerned with subjective workload analysis.

6.11.1 Time Analysis

Independent Variable	Significance Level
Orientation Update	$F(1,94) = 8.524$, $p=0.004$
Feedback Cue	$F(2,188) = 116.541$, $p<0.001$
Number of Display Elements	$F(6,564) = 5.731$, $p<0.001$
Update * Feedback Cue	$F(2,188) = 74.792$, $p<0.001$
Update * Elements Number	Not Significant
Display Elements * Feedback	Not Significant

Table 5. Between and within subjects main effects and the interaction between the independent variables in the experiment in time to select measurements. F-values and significance levels are also presented.

A (2x3x7) ANOVA with orientation update as a between subjects factor showed a significant main effect of orientation update, display type and number of display elements. The effect of the interaction between the orientation update and the other feedback cues was also significant. The results are presented in Table 5. Mean times as a function of display content for all examined cases can be found in Figure 27.

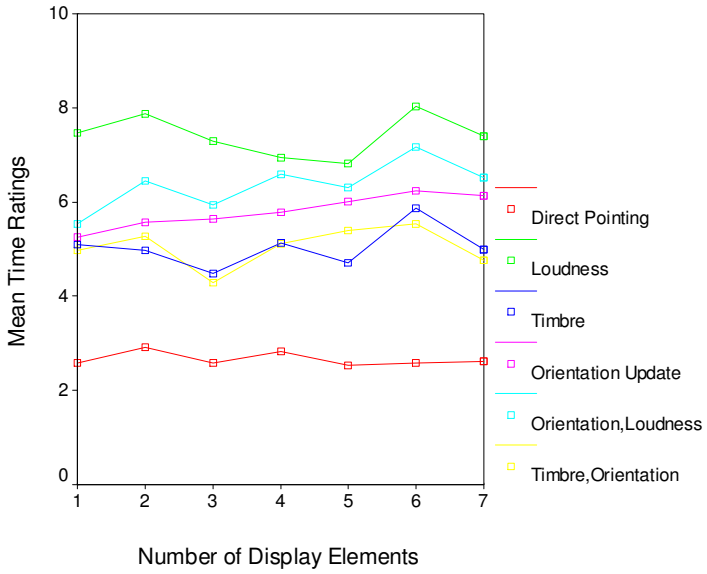


Figure 27. Mean Times to Complete Trials as a function of the number of the display elements for each feedback type used in the experiment.

Given the main effects observed, *post hoc* t-tests with Bonferroni confidence interval adjustment were performed to check for differences between the different feedback cues. All feedback cues were found to differ significantly with the exception of the two interfaces that used the timbre cue where sounds were presented one at a time. Utilization of the orientation update cue was found to slow interaction. The ordering of feedback cues with respect to speed is therefore: direct pointing, timbre, orientation update, loudness & orientation update and loudness alone (see Figure 27). Interaction using the combination of the orientation and loudness cue resulted in faster interaction compared to using the loudness cue alone but slower than using the orientation update cue. It is also interesting to observe that the interface where active listening was enabled was more sensitive to increasing the number of display elements than the interface where orientation update was disabled. The associated curves show a clear increasing trend. Table 6 presented the means and standard deviations of the time measurements.

O/U	Cue	1	2	3	4	5	6	7	Mean
Yes	None	5.3 (1.7)	5.6 (1.4)	5.6 (2.2)	5.8 (1.8)	6.0 (1.8)	6.2 (1.9)	6.1 (1.9)	5.8
	Loudness	5.5 (2.2)	6.5 (2.4)	5.9 (2.5)	6.6 (2.6)	6.3 (2.3)	7.2 (2.9)	6.5 (3.4)	6.4
	Timbre	5 (2.7)	5.3 (2.9)	4.3 (2.3)	5.1 (3.0)	5.4 (3.2)	5.5 (3.0)	4.8 (2.2)	5

No	None	2.6 (0.8)	2.9 (1.2)	2.6 (0.9)	2.8 (0.9)	2.5 (0.8)	2.6 (0.9)	2.6 (0.7)	2.65
	Loudness	7.5 (2.5)	7.9 (3.35)	7.3 (1.7)	7.0 (2.4)	6.8 (2.8)	8.0 (3.2)	7.4 (2.5)	7.4
	Timbre	5.0 (3.0)	5.0 (2.2)	4.5 (2.2)	5.1 (2.3)	4.7 (2.0)	5.9 (3.2)	5.0 (2.0)	5

Table 6. Means and standard deviations for the time measurements in this experiment as a function of display population and feedback cue. O/U refers to whether orientation update was present in the display. The measuremements are in seconds.

6.11.2 Accuracy Analysis, Throughput and Effective Target Widths

Accuracy was calculated as the absolute difference between user pointing direction and target position. A (2x3x7) ANOVA on absolute deviations from target showed a significant main effect of the orientation update cue, display type but not of number of display elements. The interaction between feedback cue and orientation update was significant, as was the interaction of display elements and the three feedback cues. Significance levels can be found in Table 7.

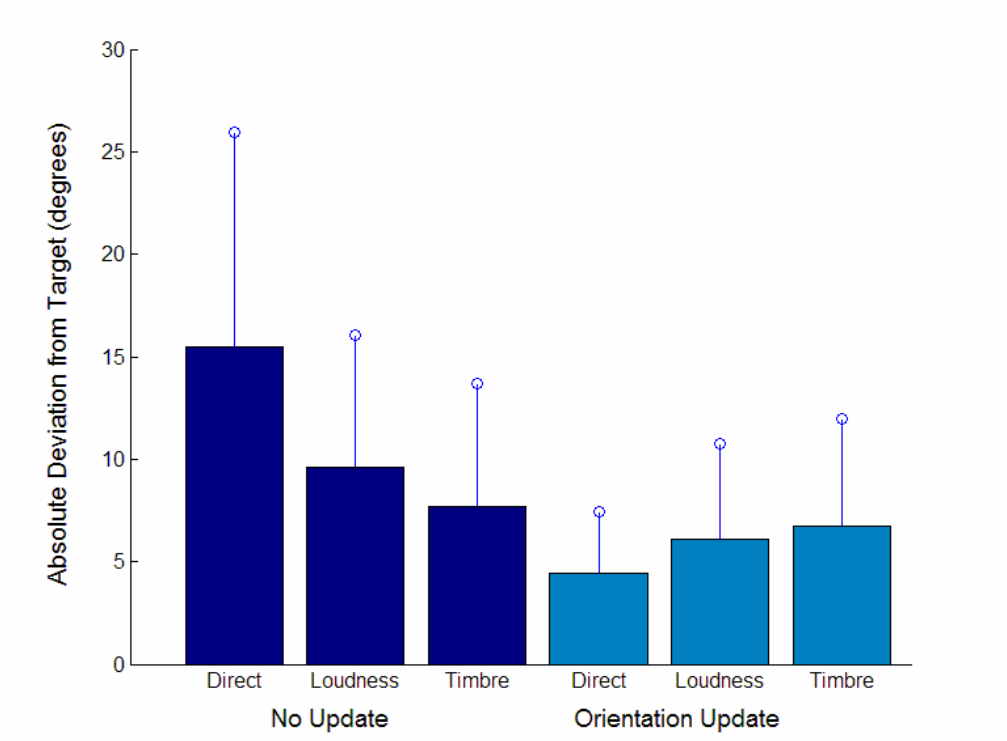


Figure 28. Mean absolute deviation from target and its standard deviation means calculated using data from all cases of the number of display elements factor.

Given the main effects observed in overall accuracy *post hoc* comparisons were performed for all combinations of feedback cues to order them with respect to accuracy. All feedback cues and combinations were found to differ significantly. Orientation update was found to significantly enhance the accuracy of selections. In general, high standard deviations were observed in user selections. The different feedback cues can be rated with respect to accuracy as: orientation update, loudness and orientation update, timbre, loudness only and finally direct pointing. Mean accuracy ratings for all feedback combinations across all display populations are plotted in Figure 28. To compare the different feedback cues in a more systematic way, the measures of *throughput* and *effective target width* are used. The accuracy ratings in our study exhibited a large amount of between-subject variation. This is due to the influence of throughput and effective width results. To give an indication, the range of throughput and effective width values for the participants of the experiment is provided in Table 8.

The between subject variation can be explained by the skill required by the tasks and the absence of any training. The data presented were measured from participants that had no experience in the sound localization task or in the use of virtual audio feedback cues. It is expected however that performance will improve through training. The results presented are therefore representative of an untrained population and thus represent a safe approach to design when using the feedback cues under study.

Independent Variable	Significance Level
Orientation Update	$F(1,94) = 854.725$, $p<0.001$
Feedback Cue	$F(2,188) = 43.552$, $p<0.001$
Number of Display Elements	Not Significant
Update * Feedback Cue	$F(2,188) = 36.674$, $p<0.001$
Update * Elements Number	Not Significant
Display Elements * Feedback	$F(12,1128) = 3$, $p<0.001$

Table 7. The effect of the independent variables on accuracy scores, F values and significance levels.

Effective target widths for the different feedback types averaged across all subjects and all display populations are presented in Figure 29. The effective target widths when doubled will result in close to perfect selection rates. In terms of target size alone, the different feedback cues are ordered as: orientation update, loudness & orientation, timbre, loudness alone and direct pointing to a static audio target. Throughputs were calculated according to Equation **28** and are presented in Figure 30. It can be observed that, despite the rather large effective target widths, direct selection proved to be the most efficient when throughput is concerned.

Cue	Throughput	Width
Direct pointing	0.55-1.33	22 – 54
Loudness	0.29-0.53	21 – 28
Timbre	0.4-0.95	14 – 28
Orientation update	0.51-0.78	8-12
Loudness & Orientation	0.38-0.94	12-19
Timbre & Orientation	0.4-1.18	13-25

Table 8. Throughput and effective target width variations between participants.

The rating of the different feedback cues in terms of throughput is: direct pointing, orientation update, timbre, loudness and orientation, loudness alone. Throughput comparisons are quite useful since they combine accuracy and timing information in one uniform measure. An ANOVA was performed on throughput and target width measures as these were obtained for each participant and for all display populations in the experiment. The result showed that display population was not a significant factor, for this reason throughput and effective width data are presented averaged across all display element populations.

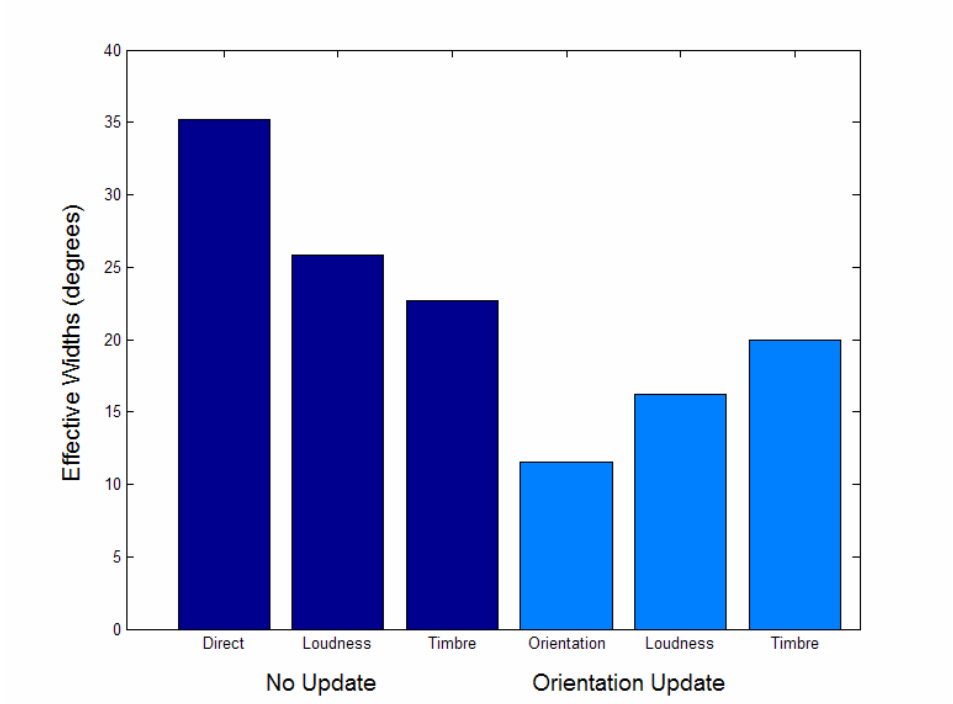


Figure 29. Effective target widths for the different feedback cues calculated across all display populations.

6.11.3 Workload Analysis

NASA TLX sheets were filled by the participants after completing each trial set. Mean workload values are presented in Table 9.

	Simultaneous	Envelope	Timbre
Exocentric	4.87 (1.09)	4.79 (1.34)	4.3 (1.58)
Egocentric	3.04 (1.00)	4.15 (1.07)	3.45 (0.92)

Table 9. Mean Workload Values for the participants that took part in the experiment. In parentheses, standard deviation values are presented.

A 2x3 ANOVA showed no significant main effect of neither feedback cues nor orientation update. It is however, worth noting that workload values between the egocentric and exocentric displays differ substantially in the simultaneous case as a t-test would reveal ($t = 3.004$, $p = 0.013$). Detailed investigation on each parameter of the NASA TLX tests showed that the orientation update cue had a significant main effect on mental demand $F(1,10) = 8.683$, $p<0.013$, time pressure $F(1,10) = 6.503$, $p = 0.027$ and frustration experienced $F(1,10) = 5.970$, $p<0.033$. When orientation update was enabled participants felt higher time pressure and characterized the task as requiring more mental demand and causing more frustration. The feedback cue had a significant main effect for the measures of mental demand $F(2,20) = 6.920$, $p = 0.005$, performance level achieved $F(2.20) = 5.825$, $p = 0.009$ and annoyance experienced $F(2,20) = 3.845$, $p = 0.037$. The silent interface was found to require less mental demand from the loudness enabled interface, it was less annoying than the envelope interface and made participants believe that they performed better than in the other cases.

6.12 Discussion

One of the major findings of this study is the time/accuracy trade-off that is associated with movement-coupled cues. Both in the orientation update and the loudness cases, participants were significantly slower than in the cases where cues were not updated continuously in the display. In such cases, although there is an improvement in accuracy, the high timing demands compromise the benefit. This is quite evident in the throughput ratings where the accuracy superiority of the movement-coupled cues was cancelled out by the increased movement times. The reason for the increased time demand is that such cues require continuous target attainment due to the fact that each movement users make affects the perceived soundscape and forces them to re-evaluate their current position with respect to the target. This can result in increased time taken especially when users are close to their final position. The demand for continuous target attainment would be relaxed by providing discrete levels for certain distance to target intervals. Increased final positioning times have been associated with orientation update in other studies, such as [68]. On the other hand, movement-coupled cues were shown to be useful in reducing

target size, as is shown by the effective target widths associated with these tasks. The most successful case of the orientation update feedback cue required just one third of the effective width required by direct pointing.

The loudness cue rated reasonably well when used in combination with orientation update, however when used alone was less effective. In this sense, the loudness cue is not very useful in assisting pointing based interactions. However, due to the fact this cue is effectively adjusting focused element to distracter element level ratios, it can be quite useful in assisting selective and divided attention in the display, an option that can be very useful when in mobile settings. The timbre related cue, rated quite well in terms of time and accuracy. However, its success was limited due to the lack of continuous contact with the target. This type of cue resulted in a searching action that reduced interaction speed. In terms of throughput, this type of cue was competitive with the orientation update cue.

When considering egocentric vs. exocentric interfaces it was found that they are both prone to interaction efficiency and effectiveness problems. In the exocentric display participants were almost three times more accurate than in the egocentric, at the same time being two times slower. The timbre cue reduced the localization error by two times in the egocentric display but it increased interaction speed by the same factor. However, it should be noted that the target sound was not audible when the timbre cue was used in this sense the effect on interaction speed is expected to be less when the user is having direct contact with the display elements as is found in Chapter 8. In the exocentric display the timbre cue only improved interaction speed by 20% but it degraded accuracy by about 70%. In this sense the improvement is not that pronounced in this case. Again, when participants were tested with the timbre cue, the target sound was not audible in the beginning of the trial. In the opposite case it is expected that the improvement on interaction speed would be higher. In addition, it should be noted that the target was in the same position in all trials. When target position is unknown it is expected that the benefit on interaction speed will be higher, however this is an issue whose clarification requires further experimentation.

It is worth considering the results of this experiment with respect to designing a spatial audio pointing task. The results show that the different cues alone are not sufficient due to the time-accuracy trade-offs that were observed.

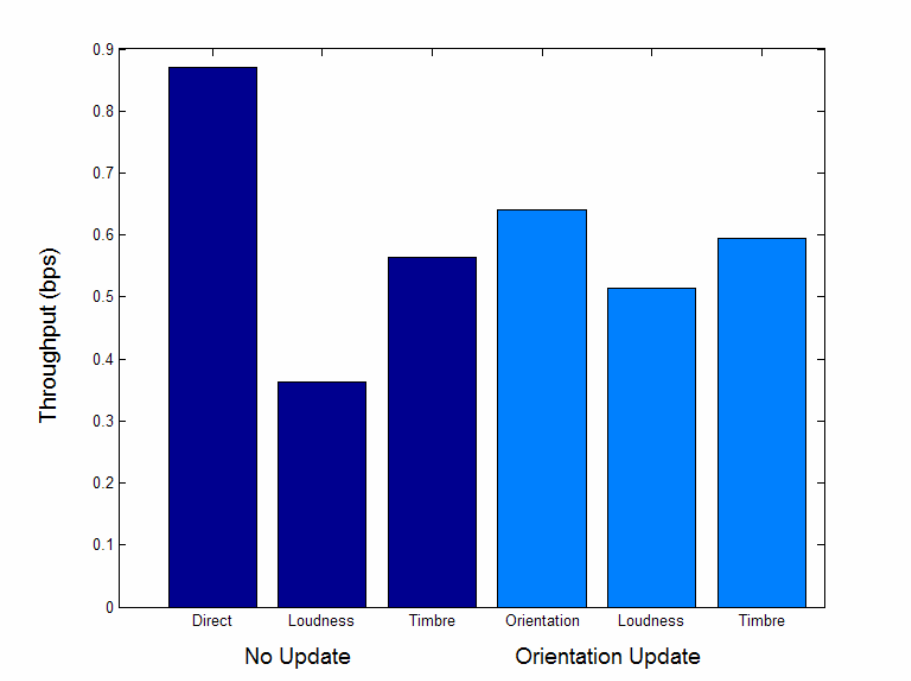


Figure 30. Mean Throughput for the different feedback cues calculated across all display populations.

Solutions to this problem can be sought in cue combinations. As has been found, performance ranges between the limits set by the individual cues when they are combined. For example, combining the orientation update cue (which was more accurate than the loudness cue) with the loudness cue resulted in more accurate performance than the loudness cue alone but was less accurate than the orientation cue alone. A similar trend was observed for the rest of the cue combinations. Combining cues results in a compromise, which can be also used constructively to enhance cues that need to be used but are lacking in a certain interaction aspect. For a task that will be performed repeatedly, combining direct pointing with a timbre cue would result in an interaction that is fast and accurate.

The experiment also focused on the effect of distracting sounds on interaction. Interaction using the non-movement coupled cues was not affected by increasing the number of distracting sounds. This is the case in direct pointing and in the timbre cue, where time to select and accuracy of selection was not affected by increasing the number of sounds in the display. For the rest, a rising trend on movement time was observed, however with no significant effect on accuracy. Time to select ranged between 5 sec. for one target and up to 6 sec for 6 or 7 elements in the display in the orientation update case, an increase of 20%. In the loudness case, a similar trend was observed.

The results of this study can be used to design improved spatial audio window applications. They can be useful in predicting performance when using a certain cue in the display and deciding on possible combinations of cues. Depending on the requirements of an application, a designer might use the results

to decide on the use of cues that can be effective in terms of effectiveness and efficiency. Due to the mobile aspect of this type of interaction this study can help in the design of usable mobile applications that take advantage of the audio modality and gesture recognition to facilitate interaction and overcome the problems that stem from the variability imposed by movement.

6.13 Conclusions, Guidelines & Future Directions

Experimental hypothesis 1 was found to be valid. Interaction was significantly slower in the egocentric compared to exocentric display. Experimental hypothesis 2 was also found to be valid, the timbre cue was found to result in usable interaction, however its efficiency was compromised by the fact that there was no direct contact to the display elements. Experimental hypothesis 3 was also verified, the loudness cue improved accuracy compared to direct pointing but it was not as successful as the orientation update cue. Experimental hypothesis 4 was found to be partly valid. It affected interaction in the exocentric display but not in the egocentric.

A study comparing feedback cues with the objective of enhancing pointing efficiency in deictic spatial audio displays was presented. Participants were tested in a systematically varied display environment to examine the effect of distracter display elements on interaction. Movement-coupled feedback cues effectively reduced effective target widths, but the efficiency of the cues was found to be compromised due to the reduction in speed caused by the requirement of continuous target attainment these cues impose. Movement-coupled cues were also found to be sensitive to display population, direct pointing cues not being affected significantly. Feedback cue combinations were found to improve the less effective cues but degrade the more effective ones. Lack of continuous contact with the target was found to negatively influence interaction speed. The results reveal that spatial audio display design is challenging, but with appropriate design it is possible to overcome interaction uncertainty and provide solutions that are applicable in human computer interaction.

With respect to the Research Questions outlined, the evaluation method used in this Chapter was found to successfully identify various interaction aspects thus paving the way for a more complete answer to Research Question 3. With respect to Research Question 1, it was found that feedback marked display areas are the best way to support spatial audio target acquisition. With respect to Research Question 2, it was found that the type of feedback used to indicate target sound position affects spatial audio target acquisition. When the feedback cue is continuously mapped to user movement, there is also an effect of distracting display elements.

With respect to the future directions, the results of this experiment provide sufficient knowledge in order to design a spatial audio target acquisition task, attempted in Chapter 7. Based on the results a number of guidelines are also provided.

- Interaction in both egocentric and exocentric interfaces without feedback is prone to effectiveness or efficiency deficiencies and the designer should seek a way to compensate for them. Feedback marked audio display areas can successfully alleviate this problem.
- A designer should not rely on a loudness based cue to deliver directional information. Such a cue is useful in improving intelligibility rates however, its success as a directional cue is limited.
- Increasing the number of display elements does not significantly influence interaction in familiar egocentric displays. In exocentric displays, however, because of the higher cognitive load imposed by the real time updated localization cues, increasing the number of display elements affects interaction speed in a negative way.

7 An investigation on the effects of Reproduction Equipment, Target Size and Inter-Target Separation on Gesture Interaction with a Spatial Audio Display

7.1 Introduction

In this Chapter a spatial audio target acquisition task is designed based on the results of the experiment presented in Chapter 6 and according to the requirements of Research Question 1. Then the task is evaluated so that inferences can be made on the effect of reproduction equipment and display segmentation. An investigation on the appropriateness of the models of visual target acquisition on spatial audio target acquisition is performed to further explore Research Question 3. In addition, three different reproduction techniques and two display segmentation strategies are examined to identify more factors affecting spatial audio target acquisition and thus providing more insight into answering Research Question 2.

7.2 Rationale

The results of the study presented in Chapter 6 showed that spatial audio target acquisition in the absence of feedback in both egocentric and exocentric interfaces is prone to either efficiency or effectiveness deficiencies. Loudness cues, although useful as intelligibility feedback, were not found to substantially improve the effectiveness and efficiency of interaction. However, the simple recognition based timbre cue was found to work in a satisfactory way both in the cases of egocentric and exocentric interfaces. Interaction using this cue resulted in the best performance both in terms of efficiency and effectiveness.

The spatial audio acquisition task will take place in an egocentric display. This decision is made due to the mobile nature of interaction that the task is expected to support. As mentioned in Section 3.3.2, egocentric designs are expected to provide better support for interactions that are of a repetitive nature, such as deictic ones, and are expected to be performed when mobile. This is due to the fact that interaction in an egocentric display is less demanding in terms of time taken and also because display element position remains unchanged relevant to the user as he or she moves in the real world.

The proposed target acquisition task that will be examined in the rest of thesis is therefore the one of selecting an egocentric spatial audio element confined in space in front of the user whose position is marked by audio feedback in pre-defined target display area that has been allocated to it.

Figure 31 illustrates the spatial audio target acquisition task that is proposed and is going to be examined in this and the next chapter. As can be seen in the Figure, the relevant parameters are these of the target area and distance to target.

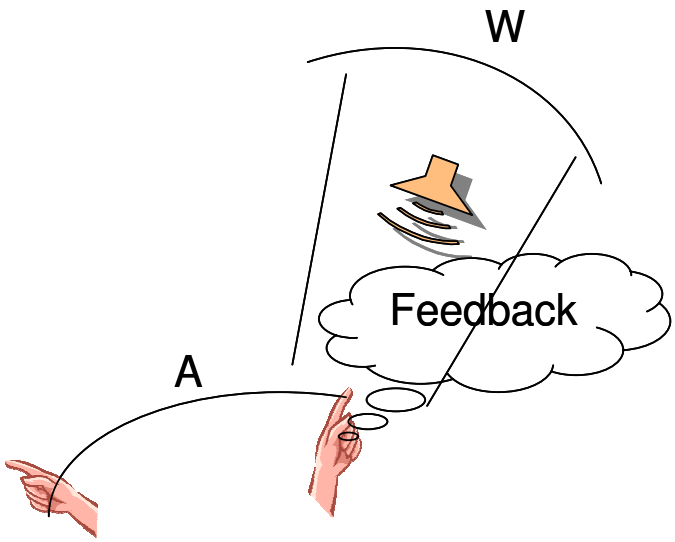


Figure 31.The spatial audio acquisition task under examination. A is the distance to target and W the target width.

The task will be examined in this experiment in two ways. The first is against three different reproduction techniques and the second is against two display designs.

7.3 Background on the examined reproduction techniques

In most of the designs, evaluations and psychoacoustic studies in the literature, both in Chapter 2 and Chapter 3, sounds in the spatial audio display are presented through headphones. Headphone presentation can, however, be a disadvantage in some application areas for spatial audio displays. Our auditory sense is valuable when mobile both for communicating and as an alerting mechanism. Blocking it can be irritating and possibly dangerous, depending on the interaction context. For example, being able to hear cars when crossing the road is important to avoid accidents. One way to overcome this problem is by using alternative reproduction devices. Nomadic Radio [103], a spatial audio interface targeted primarily at messaging, was designed to work on shoulder mounted speakers to overcome this problem. This approach, however, can be indiscrete due to other people overhearing sounds emitting from the loudspeakers. In noisy environments intelligibility is also likely to be reduced by the interference of other sounds. Goose and Safia [46] used speakers for spatial audio presentation since their proposed system

was designed for inside a car. This option however, is a context specific solution and cannot be applied generally in mobile situations.

In this experiment two alternative reproduction devices are evaluated and compared to standard headphone presentation. In particular, monaural presentation and bone conductance presentation are evaluated and compared to presentation using normal headphones. Both of these options are interesting since they provide the possibility of unblocking the audio channel at the same time as the user participates in a digital audio experience.

Monaural presentation is achieved by playing the sound using only a single earphone (such as most mobile phone speakers or hands-free kits). This technique has the advantage that it allows for one of the two ears to monitor the real audio environment. However, sound localization is based on binaural cues, i.e. differences between the signals arriving at both ears, so the spatial impression is degraded in monaural situations. The impression of space in monaural presentation does occur (mainly due to the effect of the outer ear) but localization judgments are far from accurate [14]. Therefore, it is necessary to investigate whether the localization cues are strong enough to make a successful spatial audio interface.

Reproduction using bone conductance headphones is accomplished by transmitting vibrations through the skull of the user. Such headphones feature a vibrating surface that is mounted on the side of the head in front of each ear. The mounting mechanism is very similar to standard headphones, with the difference that the outer ear is completely open. Vibrations propagate through the skull to stimulate the ear and thus become audible. The perceived sound signal will, however, be distorted by the transmition path, increasing the signal to noise ratio. Reproduction fidelity is thus lower than normal headphones. Nevertheless, this reproduction technique can lead to intelligible impressions of speech or music. This may not be the case for spatial audio, due to the fact that the subtle pinnae effects applied through HRTF filtering will be distorted. Some cues will remain, in particular inter-aural intensity and time differences. These cues can produce a spatial impression similar to stereo reproduction which may be enough to form an overview of the spatial structure of a simple audio display, for example one that is based on the horizontal axis in front of the user.

Given the disadvantages of these reproduction devices over standard headphones, an experimental evaluation is necessary to see how interaction will be affected. Studying these alternative reproduction devices is useful since it can provide insight on how to combine the real with a digital audio environment, as well as into how feedback can be used to facilitate interaction in the presence of weak localization cues.

Figure 32. The different headphone types used in the experiment, bottom left is the single earpiece (Panasonic RP-HS50), top left are the bone conductance headphones (Vonia EZ – 3200P) and to the right are the Sennheiser HD 200 headphones.

7.4 Background on the examined display segmentation techniques

From a display segmentation point of view, the question being investigated in this experiment is how to allocate display area to display elements in a uniform manner. Such a decision directly affects target size and target separation and therefore interaction. The effect of these parameters on interaction has to be taken into account before a decision on appropriate values can be made. With reference to Chapter 4, target size and distance to target affect the time to select a target and therefore interaction efficiency. The results from Chapter 5 indicate that target width affects the accuracy of user selections and therefore interaction effectiveness. However, increased target width leads to increased inter-element distance. An understanding of the effect of these two variables on interaction, both independently and relative to each other, is therefore necessary to proceed with display design in a formal way.

Target width should be adjusted depending on the gesture that is used in the display, as has been shown in Chapter 5. Obviously its size should not drop below a certain minimum value that would enable effective interaction. According to the results in Chapter 5 and Chapter 6, for a hand based physical pointing task in an exocentric spatial audio display in the presence of feedback, 16 degrees target width would result in approximately 70.7% success. Assuming a slope of one for the associated psychometric function, about 20° would be required for a selection rate in the vicinity of 100%. It should be noted that

no data on the psychometric function of this particular task have been found and thus this assumption on the slope of the function is purely hypothetical. However, it is used later on in the study and for this reason it is derived here.

When seeking a way to characterize the effect of target size and distance to target on spatial audio target acquisition, it is unavoidable to draw on existing results in visual target acquisition that were presented in Chapter 4. Selecting a feedback marked audio display element based on the direction of the sound event either by a real or virtual pointing gesture has many similarities to homing to a visual target, as in Fitts' law experiments. This is due to the fact that participants are informed on the direction of the sound event by spatial audio and they are assured on the target boundaries by imposed audio feedback. However, a different sensory modality is used for event localization compared to vision and as has been described in the previous section, users have a less precise impression of target location. Results of studies focusing on visual targets can therefore serve well as a starting point that can help to identify parameters that affect this type of interaction. For this reason further analysis in the thesis is based on the quantities of distance to target and target width since they have been proven to affect virtually all pointing tasks and serve as a well founded starting point for such an investigation. It is hypothesized that interaction in a spatial audio display is affected by the prominent variables of target width and distance to target in a manner similar to what has been described by Fitts' law [73]. However, given the novelty of the interaction technique and the absence of any results in the literature, the pointing based 3D audio selection task has to be examined to reach conclusions on the properties of this interaction technique.

As discussed in Chapter 4, all relevant models of visual target acquisition are a function of distance to target over target width. Display design choices such as target size and inter-target separation depend on the relative salience of distance and width as well as on whether the actual model is logarithmic or linear. If the saliency of distance to target is equally important to target width then any increase in distance followed by an equal increment in target width will not affect time to select the target. If distance has more salience than width then display elements must be placed as close as possible keeping a reasonable target width. On the other hand, if width prevails then it is worth placing the sounds in the display utilizing the whole display area.

In an attempt to make an initial investigation on the aforementioned issues interaction is examined in two display settings, one where sounds are placed as close to each other as possible and one where sounds are placed as far as part from each other as possible, while maintaining a constant A/W (distance to target to target width) ratio in both displays. In order to test which of the hypotheses prevails, two displays were designed that are characterized by equal ratios of distance to target and target width. The MINIMAL interface contained four sounds each having target width of 20°. Sounds were placed every 20° starting from -30° and up to 30°. Sound locations were thus at -30°, -10°, 10°, 30°. The MAXIMAL interface also contained four sounds, each having target width of 45°. Sounds were placed every 45° starting from -67° and up to 67°. The interface was placed on the circumference of a circle in the

horizontal plane with 0° in front of the user's nose. Experiment trials were designed to require participants to move between the available position pairs in both interfaces thus resulting in distance arcs of 20°, 40° and 60° for the MINIMAL interface and 45°, 90°, 135° for the MAXIMAL interface. The ratios of distance to target to target width were constant in both displays having values of 1, 2 and 3.

7.5 Experimental Design

The experiment design is presented in Table 10. Participants were split in two groups. Variable Display was tested as a within-subjects variable all participants being tested in both displays in a counterbalanced order. Variable Reproduction was tested as a between subjects variable to reduce experiment time. Half of the participants performed the experimental task wearing normal and Bone Conductance headphones, and the other half wearing normal headphones and monaural presentation. Distance to target and target width values for the two displays as well as the target locations are presented in Table 10. These variables are associated to the variable Display however they will become relevant in the experiment task as well as in the Results Section.

	Reproduction	Display	A's / W	Target Locations
Group A	HD	MINIMAL	20°, 40°, 60°/20°	-30°,-10°,10°,30°
		MAXIMAL	45°,90°,135°/45°	-67°, -22°, 22°, 67°
	BC	MINIMAL	20°, 40°, 60°/20°	-30°,-10°,10°,30°
		MAXIMAL	45°,90°,135°/45°	-67°, -22°, 22°, 67°
Group B	HD	MINIMAL	20°, 40°, 60°/20°	-30°,-10°,10°,30°
		MAXIMAL	45°,90°,135°/45°	-67°, -22°, 22°, 67°
	MA	MINIMAL	20°, 40°, 60°/20°	-30°,-10°,10°,30°
		MAXIMAL	45°,90°,135°/45°	-67°, -22°, 22°, 67°

Table 10. Experiment Design: HD stands for Headphones, BC for Bone Conductance, MA for Monaural, A for distance to target and W for target width. The independent variables reproduction and display type are shown together with their associated values and the design choices they had resulted into for the sound positions in the display, their target width and the distances between them.

7.5.1 Experimental Hypotheses

Prior to the experiment it was hypothesized that:

1 Reproduction Technique will affect time to select the target. Interaction speed is expected to decrease in the bone conductance and monaural cases, because of the weaker localization cues

2 Accuracy is not expected to be affected by reproduction technique because of the feedback marked target areas that are considered sufficient to guide users to the target.

3 With respect to the effect of the two display arrangements, Fitts' law would predict a similar time to select the target.

4 An improvement in accuracy would be expected by the larger target sizes in the MAXIMAL display, however since 20° are considered to result in selection success rates close to the performance ceiling the extent of the improvement should not be very big.

7.6 Experimental Task

The experimental task was designed to represent a common scenario in interaction with a deictic spatial audio display where users must select a sound emitting from somewhere in space in front of them using a tracker held in their hand. Participants initially had to listen for a target sound, played in isolation from a certain position in space. The target sound occurred in one out of the four display element positions. The remaining three display element positions were filled with distracter sounds. In each trial an announcement of target sound position was done and then the target sound played continuously together with three distracter sounds. To select the target sound participants had to point at its location and make a downwards wrist gesture to indicate selection. Participants received audio feedback (the sound of people cheering) when they were within the target sound's area. Sounds were placed according to the MAXIMAL or the MINIMAL specification (see Table 10) in the horizontal plane. The target sound was placed in the leftmost position in every second task to equalize the distance pairs.

7.7 Stimuli & Apparatus

The aim of the experiment was to assess the effects of the three different reproduction devices on target selection performance in a spatial audio interface. Figure 32 shows the different devices used. These were: standard Sennheiser HD250 closed-back headphones, Vonia EZ-3200P bone conductance headphones and a Panasonic RP-HS50 earpiece (reproduction using standard headphones will also be referred to as binaural listening). An XSENS MT-9B orientation tracker (www.xsens.com) was used to track user orientation and the selection gesture. DieselPower Engine (www.am3d.com) was used to spatialise the sounds in the display.

To avoid any effects related to timbre, the same sound was used for both distracters and the target sound. This was a short (0.5 sec.) segment of white noise. To improve intelligibility, a 300ms onset difference between neighbouring sounds was introduced. Counting from left to right, this resulted in the second sound starting 300ms later than the first sound, the third 600 ms later and the fourth 900 ms later than the first sound. Sounds repeated after a 500 ms period of silence. All sounds in the display were played at the same level.

7.8 Procedure & Participants

Participants performed the task according to the design presented in Table 10. There was a short training session prior to testing, during which their performance was monitored to make sure that they understood the task. After participants successfully completed four consecutive trials during the training session, the testing started. Participants were tested in the two experimental conditions associated with the groups shown in Table 1, one followed by the other in a counterbalanced order.

Sixteen participants were tested (17 to 27 years of age, mean age of 21 years, 2 females and 14 males). Participants were paid £5 for their participation. Participants were asked whether they were facing any hearing deficiencies and if they did they were excluded from the experiment. Time to complete trials, angular deviation from target and movement pattern was recorded for each trial during the experiment. After testing in each experimental condition participants completed a modified NASA TLX subjective workload assessment form that included the factor of annoyance [19, 54]. The experiment lasted half an hour. In total 128 measurements were available per level combination for the monaural and bone conductance cases and 256 for the standard headphone case. The setup was the same as in Chapter 6. Figure 26 can be used as a reference.

7.9 Results

The analysis presented here is based on the raw data that are available in the associated section of Appendix 1. The analysis involved two within-subjects comparisons, one for each participant group and one between-subjects comparison to compare bone conductance and monaural presentation.

7.9.1 Time Analysis

The time taken to complete trials was analyzed using a repeated measures ANOVA. Within subjects ANOVAs were used to test monaural presentation vs. headphone presentation and bone conductance vs. headphone presentation and a between subjects ANOVA was used to test bone conductance vs. monaural presentation. The aforementioned analyses followed a 2 (reproduction technique) x 2 (display type) x 3 (A/W ratio) design.

In all analyses display type was found to have a significant main effect on interaction $F(1, 127) = 74.214$, $p<0.001$ in the headphone vs. monaural comparison, $F(1,127) = 136.23$, $p<0.001$ in the headphone vs. bone conductance comparison and $F(1,254) = 109.282$, $p<0.001$ in the bone conductance vs. monaural comparison. As is also evident from Figure 33, interaction was significantly faster in the MAXIMAL interface.

With respect to the effect of reproduction equipment on interaction speed, no significant difference was found between the headphone and bone conductance presentation. However, there was a significant

main effect of reproduction technique between the monaural and headphone conditions $F(1,127) = 166.234$, $p<0.001$ and the monaural and bone conductance conditions $F(1,254) = 2769.391$, $p<0.001$.

A/W ratio was also found to have a significant main effect in all comparisons, $F(2,254) = 9.152$, $p<0.001$ in the headphone vs. monaural comparison, $F(2,254) = 35.687$, $p<0.001$ in the headphone vs. bone conductance comparison and $F(2,508) = 7.659$, $p=0.001$ in the bone conductance vs. monaural comparison. The effect of A/W however, varied uniformly only in the headphone and bone conductance case of the MAXIMAL display.

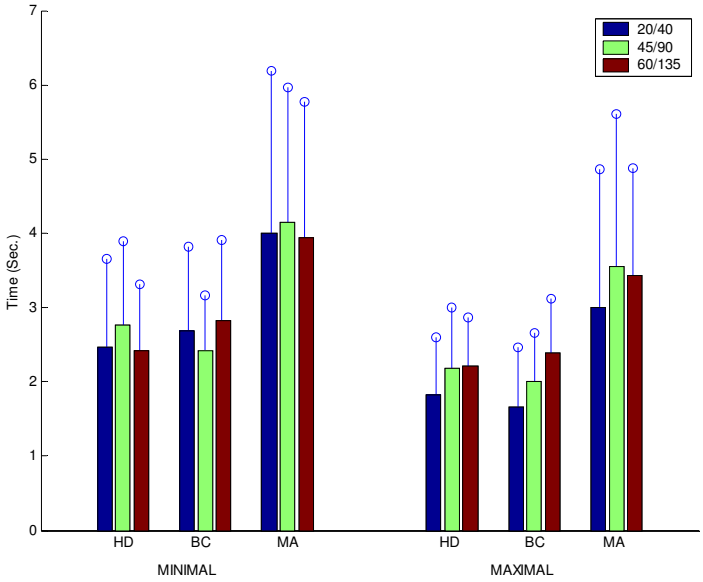


Figure 33 Mean time and standard deviation for the three presentation methods the two display designs and associated distance paths. The first line in the legend corresponds to the distances in the minimal display, while the second in the distances in the maximal display.

7.9.2 Accuracy & Workload Analysis

Based on the orientation measurements for each selection, percentage correct ratings for each condition were calculated and can be found in Table 2. As can be observed the wide angle span of each target resulted in high success rates for all reproduction types. Reproduction device was not found to affect accuracy.

	MINIMAL				MAXIMAL			
	1	2	3	Mean	1	2	3	Mean
HD	94.53 (0.06)	92.18 (0.07)	87.1 (0.09)	91.27	99.6 (0.02)	94.14 (0.09)	93.75 (0.08)	95.83
BC	92.97 (0.05)	85.16 (0.14)	88.28 (0.08)	88.57	100 (0)	93.75 (0.09)	94.53 (0.07)	96.09
MA	92.97 (0.06)	96.09 (0.07)	92.97 (0.07)	94.01	97.66 (0.03)	98.4 (0.03)	93.75 (0.07)	96.6

Table 11. Accuracy Success Rates (%) for all Reproduction Techniques, Displays and A/W Ratios. In parentheses standard deviations are given. (HD stands for Headphones, BC for Bone Conductance and MA for Monaural presentation)

Three ANOVAs were performed on the accuracy scores as in the Time analysis. Reproduction Technique did not have a significant main effect on success rates when comparing bone-conductance and monaural reproduction or when comparing bone conductance and headphone presentation. There was a significant main effect of reproduction technique when comparing headphones and monaural presentation, $F(1,7) = 10.188$, $p = 0.015$. By inspecting Table 11, it can be inferred that this difference is localized in the MINIMAL case, where participants were on average more accurate in the monaural case than in the headphone presentation case.

Display type had a significant main effect on success rates when comparing bone conductance with headphone presentation $F(1,7) = 19.723$, $p = 0.003$ as well as when comparing bone conductance with monaural presentation, $F(1,14) = 6.463$, $p = 0.023$. There was no significant main effect of display type when comparing headphone and monaural presentation.

Finally, there was a significant main effect of A/W ratio when comparing bone conductance vs. headphone presentation $F(2,14) = 5.408$, $p=0.018$, when comparing monaural vs. headphone presentation $F(2,14) = 4.829$, $p = 0.025$ but not when comparing monaural vs. bone conductance presentation, $F(2,28) = 2.568$, $p = 0.095$.

Data from modified NASA TLX forms (including an annoyance factor) that measure subjective workload were also analyzed. No significant difference in overall workload was found in any of the comparisons of reproduction equipment. Therefore, subjective workload was not affected by reproduction type. The effect of display type on subjective workload could not be identified due to the experimental procedure. Participants were tested in both display types in a continuous way with no break and were unaware of the change in the display type during the experiment.

7.9.3 Additional Observations

Movement trajectories were analyzed and the mean time participants spent in overshooting the target during all trials and for all sound positions in the experiment were calculated. Figure 34 shows that participants spent more time overshooting in the MINIMAL display arrangement, compared to the MAXIMAL, as would be expected. In particular, the 10° sound location resulted in the highest overshooting time. The MAXIMAL interface resulted in negligible overshooting time for most of the cases, as the targets were large.

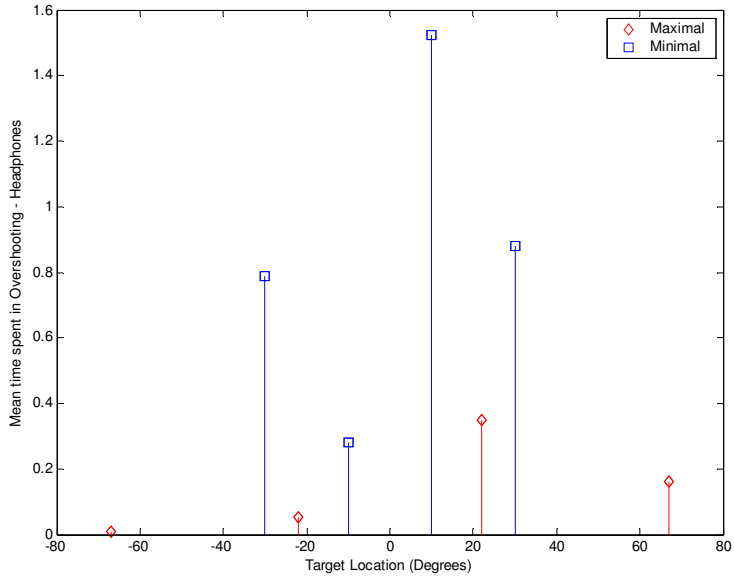


Figure 34. Mean time spent in overshooting the target per display type, headphone case.

The same procedure was repeated for the bone conductance headphones, mean time spent in overshooting the target is presented in Figure 35. It can be observed that in both cases participants frequently overshoot the target when interacting in the MINIMAL display.

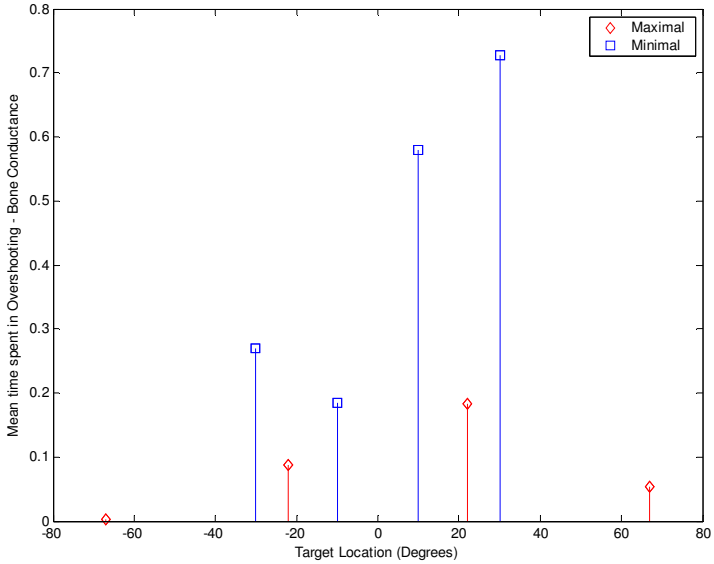


Figure 35. Mean time spent in overshooting the target in the case of bone conductance headphones

In addition, histograms were calculated, for the two display types with respect to the position at which participants indicated selection. Figure 36 shows selection histograms for the headphone case.

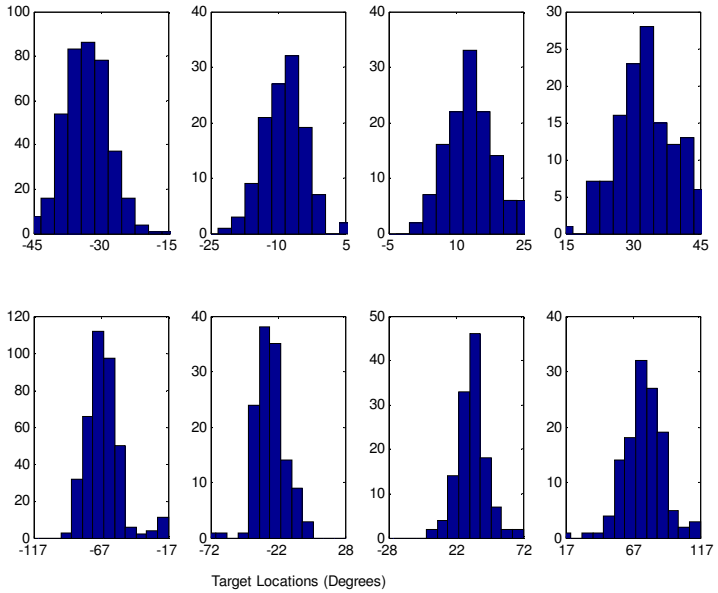


Figure 36. Histograms of selection angles for all target positions, headphone presentation.

Histograms for the bone conductance and monaural cases are presented in Figure 37 and Figure 38.

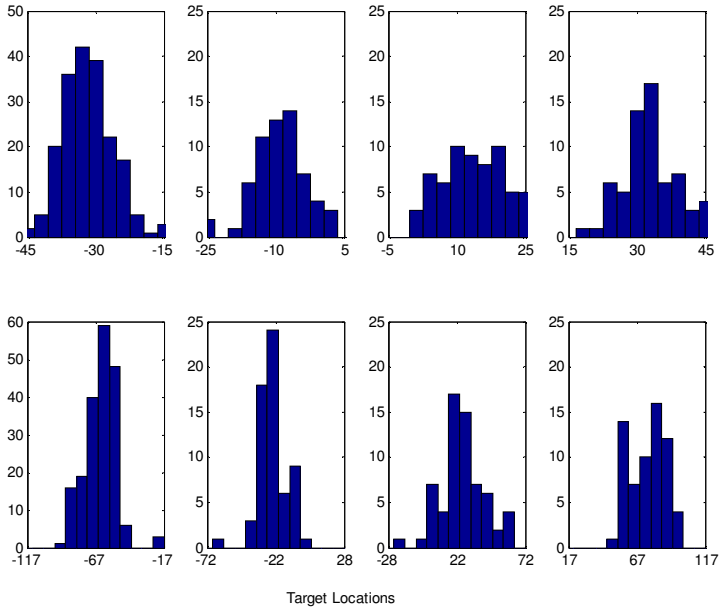


Figure 37. Histograms of selection angles for all target positions, bone conductance presentation

Participants were relatively consistent in their selections and they targeted in a manner that is quite close to the normal distribution. It is also interesting to see that the most frequent selection angles were quite close to the actual sound positions. However, in most of the sound locations, participants were selecting slightly after the target position.

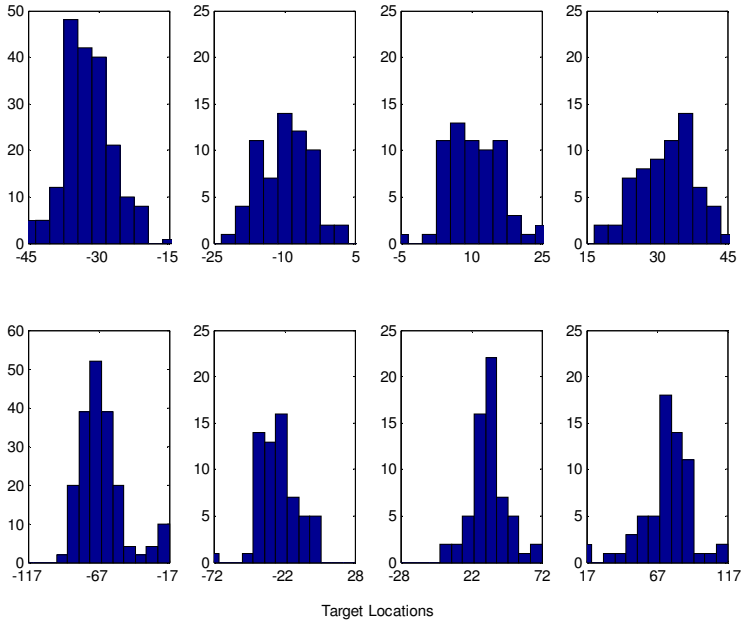


Figure 38. Histograms of selection angles for all target positions, monaural presentation

A one sample Kolmogorov-Smirnoff test was performed to test for normality of the distributions. In the headphone case, in all but the -67° and 30° locations, the histograms proved to follow the normal distribution. The results are summarized in Table 12. Results for the bone-conductance case are presented in Table 13. Finally, results for the monaural case are presented in Table 14.

\°	-30	-10	10	30	-67	-22	22	67
Z	1.2	.8	1.2	1.5	4.3	1.1	0.8	1.2
S	.12	.53	.13	X	x	.16	.48	.13

Table 12. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution for the headphone case.

\°	-30	-10	10	30	-67	-22	22	67
Z	1.064	.566	.788	1.094	2.382	1.173	.749	.906
S	.207	.906	.564	.182	X	.128	.629	.384

Table 13. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution for the bone-conductance case.

\°	-30	-10	10	30	-67	-22	22	67
Z	2.287	.689	1.233	.697	4.218	1.041	.439	.609
S	X	.730	.096	.716	X	.229	.991	.852

Table 14. One sample Kolmogorov-Smirnof Z scores using significance levels determined by the asymptotic distribution for the monaural case.

The time scores for the headphone and bone conductance in the MAXIMAL case appear to follow a modelling option that might conform to the models presented in Chapter 4. In particular linear regression is performed on mean selection times with respect to Equation 6 and Equation 8 to find out whether the results can be modelled by any of them.

In the MAXIMAL display headphone case, the logarithmic model results in an intercept a = 1.45 and slope b = 0.4044, and the regression accounted for 94% of the variance (r^2 = 0.8887). The linear model gave an intercept value of 1.67 and a slope of 0.19 while it accounted for 90% of the variance (r^2 = 0.82). While a large amount of the variance was accounted by the models, the correlation was not significant, F = 7.99, p = 0.21 for the logarithmic model and F = 4.56, p = 0.28.

The procedure proposed by Friedlander [39] was used for comparing the goodness of fit of the two models and although Fitts' model accounted for more of the variance, the difference between the goodness of fit of the two models was not found to be significant using the Hotteling's t-test. However, based on r^2 values only it was decided to use Fitts' model to calculate a performance index that can be indicative at this early stage. The Index of Performance for the MAXIMAL case was 2.47. The data for the MINIMAL case could not be explained satisfactorily by any of the models therefore no further analysis was performed.

The same procedure was repeated for the bone conductance case. The logarithmic model resulted in an intercept value of 0.9 and a slope value of 0.72 while the linear model in an intercept value of 1.2 and a slope value of 0.37. The correlation was significant in the case of the linear model F = 1.38, $p<0.001$ with the model accounting for 100% of the variance (r^2 = 1), while the logarithmic model did not correlate significantly by a small amount, F = 63.67, p = 0.07. According to Hotteling's t-test, the linear model is a better predictor of performance in the bone conductance case, t = 4.22, $p<0.01$.

To complete the picture the standard deviations of selections are presented next. It was observed that the standard deviation of deviations from target becomes higher as the target width is increased. Participants adjusted to the larger feedback area and their selections were more spread. It is also interesting to observe that the standard deviations are slightly higher in the monaural case.

\°	-30°	-10°	10°	30°	Mean	-67°	-22°	22°	67°	Mean
Headphones	5.17°	4.68°	6.3°	8.92°	6.26°	21.38°	10.99°	15.43°	15.94°	15.94°
Bone Conductance	5.53°	5.69°	8.41°	6.98°	6.66°	16.14°	10.27°	16.33°	12.15°	13.73°
Monaural	7.99°	5.36°	8.5°	6.7°	7.14°	27.76°	12.18°	11.32°	15.85°	16.78°

Table 15. Standard deviations of selections with respect to target sound position.

7.10 Discussion

The study provides a multitude of findings that have to be seen separately in order for useful conclusions to be made. For this reason, the implications of the results related to the three difference reproduction techniques are examined first. The implications of the results related to the two display arrangements and the observations in movement and accuracy patterns and the modelling of time to acquire the target are presented in the second part of the discussion.

7.10.1 Discussion on the Effects of Reproduction Equipment

The results of this study show that alternative reproduction techniques can replace standard headphones in spatial audio systems. Interaction under the three reproduction techniques was consistent in terms of the time differences observed between the MAXIMAL and the MINIMAL display. Accuracy was high and within the same range in all display settings and participants scored high in terms of success rates. In this sense the three reproduction techniques can be considered effective, with two notable differences. The first is the lack of sufficient directional cues in the monaural case and the second the linear vs. logarithmic contrast in time to acquire the target in the MAXIMAL display as a function of distance to target to target width ratio between the bone conductance and the headphone presentation.

Monaural presentation was found to slow interaction. This is due to the fact that binaural differences important in making judgments of sound direction are not available under monaural conditions. This results in extended search times for the target display elements, a fact that slows interaction. In fact, as can be observed from the results, monaural presentation slowed interaction by one and half to two times in the context of our task. However, the time to complete tasks was not completely unrealistic from a usability point of view. This suggests that in a display where the user is familiar with the positions of the elements, interaction speed under monaural reproduction is likely to reduce. This reproduction technique could also be used for presenting speech, playing music and other content presentation tasks. In addition, our auditory system is still sensitive to loudness, pitch and other sound attribute differences under monaural conditions. These could also be used to guide the user to a target sound, compensating for the lost directional cues. However, extra care must be taken if presenting simultaneous audio streams under monaural conditions because the phenomenon of masking is much stronger in monaural cases than in binaural ones as discussed in Section 7.3. This leads to the conclusion

that under monaural conditions the amount of content that can be rendered simultaneously in the display will definitely be lower than in the binaural case.

In the context of the experimental task, bone conductance presentation was found to be as fast and in the same range of accuracy as binaural presentation. Although it cannot be argued that bone conductance headphones can produce a comparable spatial impression to standard headphones, it is the case that the more 'stereo like' cue was sufficient to guide the users to the target sound. It should also be stressed that the easily perceptible audio feedback cue contributed significantly to the success of users both in the bone-conductance and the monaural cases. Rapidly presented and perceived feedback is very important with low-fidelity spatial audio reproduction techniques, because it can compensate for the weaker localization cues. Consequently, a 'stereo' like directional cue combined with good feedback can successfully guide users to a spatially positioned target sound.

However, time to select the target varied logarithmically as a function of distance to target to target width in the headphone case while it varied linearly in the bone conductance case. The difference in the time measurements however is not significant in terms of time taken for the A/W ratios used in this study. This fact implies a different selection motor strategy in the two cases. As can be observed by inspecting Figure 34 and Figure 35, participants spent more time in overshooting in the headphone than in the bone conductance case. This implies that participants moved more slowly towards the target in the bone conductance case thus being able to stop their movement when the feedback marked area was reached. Such a behaviour could be explained as a movement with a relatively constant speed towards the target as if searching for the feedback area, a behaviour similar to the one observed in the Friedlander study [39]. In other words, participants were not very confident on target position and for this reason they proceeded with the task in a careful manner. In the headphone case however, the large time spent in overshooting reveals that participants were more confident in the target position, however, due to the localization error they often overshoot it. Such a process resulted in the necessity for more corrective submovements, which in effect resulted in the more smooth logarithmic relation between time and the ration between distance to target and target width.

7.10.2 Discussion on the Effects of Display Segmentation and Interaction Patterns

The MAXIMAL interface proved to be significantly faster than the MINIMAL in all cases of reproduction technique used, although participants had to move a longer distance to reach the target. Due to the fact that the ratios of distance to target and target width (A/W) were constant in both display settings, it can be concluded that the relative salience of target width was higher than that of distance to target for the two display arrangements in the context of the experimental task. From a design point of view this indicates that it is beneficial to allow the target width to grow when space in the display is available. This result is also justified from a psycho-acoustical point of view since increased target width results in increased target separation, a situation that is known to benefit display intelligibility. As is

discussed in [109] intelligibility for audio selective and divided attention tasks is improved with higher spatial separation of display elements. As a side finding the results of the experiment indicate that the task of the acquisition of a feedback marked audio element area is sensitive to scaling the A/W ratios, a fact that has been verified to also be the case for visual target acquisition tasks [40, 52].

In the bone conductance and headphone cases of the MAXIMAL display it was found that trial completion times depended on the ratio of distance to target over target width in an ordered way. This was in accordance with the interaction models presented in Chapter 4, which relate distance to target to selection time.

In terms of accuracy, participants were successful in both displays however, a time cost appeared in the MINIMAL display which contained narrower targets. Considering the psychometric function associated with the particular task it is possible that the 20° interval lies prior to the performance ceiling, rather in the area where improvement on selection rates is still possible. The lower selection success rate in the MINIMAL display arrangement verifies that the area of 20° was more difficult to target compared to the 45° one in the context of this particular experiment. A small degradation in accuracy was found when A/W rations were increased. Such a finding is consistent with the findings in Chapter 4. In our particular case this phenomenon can also be explained by considering the gesture involved and the tracking technology used in the experiment. Due to the fact that an orientation tracker was used, participants had to adjust the direction of their palms relative to their arms when making selections. Given that all participants were right-handed it can be hypothesized that this is more conveniently done for targets to the left. When the target is to the right it is hard to fully point to the correct direction without stretching the arm or turning the body. For this reason, when participants are required to move a large distance to reach a target or when performing a selection in a specific direction that is uncomfortable it is advisable to use higher target widths to compensate for accuracy and timing deficiencies. These effects were more evident in the case of the headphone and bone conductance reproduction rather than in the monaural case where participants were forced by the task to pay extra attention.

When considering the results of the experiment from the point of view of the interaction models presented in Chapter 4, two major discrepancies arise. The first relates to the fact that although A/W ratios were the same between the two displays significant time differences have been observed. A second discrepancy was the fact that movement times in the MINIMAL display were not consistent with the behaviour predicted by any of the interaction models. The only objective finding that can be used to explain these discrepancies is the increased overshooting time in the MINIMAL display. However, a variety of factors might have contributed to such an event and for this reason it cannot be uniquely assigned to one of them. The factors that are considered relevant are confusions with respect to target position due to the distracter sounds, the higher values of target width and distance to target in the MAXIMAL case and the apparently higher salience of target width versus distance to target and a possible side bias with respect to the target position that has also been observed in other studies.

Although the target size in the MINIMAL case was big enough to allow effective selection, it might have been narrow in the way that more time was required for the participants to place their hands with confidence inside it. This might have resulted in an interaction pattern different than the one observed in Fitts' law tasks. Participants had to concentrate more and for this reason extra time was necessary for them to place the tracking device inside the feedback area.

With respect to the effect of distracting sounds, it could be hypothesized that because the timbre of both target and distracting sounds was white noise and that the onset separation was not particularly long, participants might have confused the target with the distracter while performing the task.

Finally, as has been observed in other studies [8], spatial audio stimuli emitting in the area between directly in front and the sides tend to be perceived with a bias towards the sides. In the headphone case of the MINIMAL display where sounds were not close to the sides, such an effect might have contributed to a more hasty ballistic first movement that was directed to a target area that appeared in a location displaced more to the sides compared to the target location. In the bone conductance case however, this effect was found to be less probable due to the more careful movement strategy participants adopted. In order to clarify the reason a new task has to be devised which will not involve distracter sounds and will use more sound locations than the one used in the experiment. The time difference effect can be explained logically, given the fact that for very low target widths one can expect that when keeping A/W constant and decreasing target width at some point the very low target width value will result in significantly slower acquisition times.

It is interesting to observe the distributions of selection angles. The majority of selections followed the normal distribution and selections in only in a few of the sound locations histograms deviated from the normal. In the case of headphone and bone conductance presentation histograms were sharper compared to the monaural case, a fact that can be assigned to the more clear perception of the direction of the feedback sound. In [3], non-directional feedback on target position resulted in wider histograms of selection angles, compared to the non-feedback case, however no test on normality was presented. A clear perception of the directionality of feedback can be therefore assumed to assist in sharpening the histograms of selections. This essentially implies that the stronger the directional cue the closer selections will be to the actual sound position.

The MAXIMAL interface was consistent with the Fitts' formulation. This is verified by the regression results which could account for almost 90% of the variance. The index of performance for the particular interface was approximately 2.5. This is less compared to the performance indices found by MacKenzie in [77] for pointing using three different devices in a visual target acquisition task. MacKenzie found performance indices of 4.5, 4.9 and 3.3 for pointing to a visual target using a mouse, tablet and trackball devices respectively. It can be argued though that the presence of the distracter sounds and the use of the same sound for elements in the display negatively affected the performance index. In a simpler display the performance index would reasonably be expected to be higher.

The sound stimulus that was used in the experiment was white noise for all display elements. This is beneficial with respect to sound localization since white noise is a broadband type of stimulus and also inhibits a great deal of spectral variation over time. The above characteristics are considered to assist sound localization [14]. In a spatial audio display used to accomplish human computer interaction white noise does not provide a good solution for element sonification. Other types of stimuli are more appropriate since white noise is not suitable for delivering semantic information. For a display designed to align elements based on azimuth, it is not expected that localization performance will diminish much when other types of sound stimuli are used. In fact no confusions or other deficiencies have been observed for sounds that are constrained with respect to elevation as in the display used in this particular study. In this sense, although localization might be slightly worse depending on the sounds in the display, the overall direction of the stimuli will be recognized. This fact, in combination with feedback marked display elements will result in interaction behaviour similar to the one observed in this study. If elevation was to be used the situation might become problematic since confusions are more likely to happen. The presence of audio feedback however will certainly help alleviate this problem, search time for display elements however might become an issue if confusions are common in the display.

7.11 Conclusions, Guidelines and Future Directions

Experimental hypothesis 1 was verified, reproduction technique significantly affected time to select the target. However, there was no difference between the headphone and bone conductance cases. Experimental hypothesis 2 was verified, feedback was clearly audible in all reproduction cases and reproduction technique did not affect the accuracy of selections. Experimental hypothesis 3 was not verified, time to select was different in the MINIMAL and MAXIMAL cases due to a scaling effect that was observed. Experimental hypothesis 4 was verified. No significant improvement in accuracy was observed in the MAXIMAL display due to the fact that the 20° interval used in the MINIMAL display was enough to reach a ceiling effect on accuracy.

The results of the study showed that the spatial audio target acquisition task that was designed and examined is robust, efficient and effective and can be performed in the presence of distracting sounds even when these share the same timbre and are separated by a small time difference from the target sound. In this sense, an answer to Research Question 1 has been found. In addition the selection task was found feasible to perform using bone-conductance and monaural presentation mostly due to the positive effect of easily perceptible feedback. The success of alternative reproduction techniques is promising for overcoming the problem of user isolation from the real audio world without compromising performance. The results showed that, with appropriate design, interaction with a spatial audio display using bone conductance headphones can be as fast and accurate as interaction using standard headphones. Although monaural presentation was found to slow interaction, the selection times in our study showed that the task is not unfeasible under monaural presentation. In this sense and with respect to Research Question 2,

another factor that emerges which affects spatial audio target acquisition is the quality of the localization cues that are made available to the user.

This study examined spatial audio display design by investigating the relative salience of distance to target to target width. The results showed that target width was more important than distance to target in the context of pointing-based gesture interaction with a spatial audio display. Increased target size improved time ratings in the spatial sound selection task and was found to be a useful tool in accounting for misperceptions of sound source positions and direction incurred weaknesses of the motor mechanisms that support physical gestures. However, the considerations of Guiard [52] should be taken into account. It is the case that the experimental manipulation co-varied difficulty as well as scale. The solution would be to test at fully crossed target width and distance to target values. However, this was not possible in the scope of this experiment. In this sense the result should be taken as an indication which would require further experimentation to confirm.

The use of a spatially positioned sound as audio feedback resulted in a normal distribution accounting for the participants' selections. The results showed that, given a sufficiently large target size, spatial audio target acquisition in the presence of audio feedback can be modelled using the Fitts' formulation when the display was presented through headphones. This justifies the evaluation method inspired by visual target acquisition and provides further insight to answer Research Question 3. When bone conductance presentation was used target acquisition was significantly better modelled by a linear model, a fact that implies that participants used a more careful selection strategy, due to the deterioration in the quality of the directional cues.

The utilization of such models paves the way for direct comparison of interaction techniques for spatial audio target acquisition and modality independent comparisons of pointing based interaction techniques. The Index of Performance for gesture based spatial audio target acquisition was found to be less than but comparable to virtual pointer visual target acquisition tasks and the time and accuracy ratings support the claim that interaction with a spatial audio display can be applied in real world applications.

The study also provided novel findings with respect to Fitts' law research. It was found that target width has a greater salience than distance to target in the context of our experimental task and also that differences appear in the time to acquire the target when A/W ratio varies. In addition it was found that models of target acquisition could not apply in the case of the MINIMAL display. This finding cannot be uniquely accounted for by the study, since it is unclear whether it should be attributed to the distracting sounds, misperceptions of sound position or to the effect of small target width. Further investigation is therefore necessary to clarify this issue in a spatial audio acquisition task with no distracters involved. Such an investigation is undertaken in the next chapter.

Based on the results of the study a number of guidelines for developers have been developed:

- When keeping contact with the real world audio environment is important, designers should consider the possibility of monaural and bone conductance presentation of the display.
- Designers should be careful when scaling A/W ratios because significant differences in time to select emerge despite A/W ratios being the same.

8 An Investigation into the effects of Mobility, Feedback and Index of Difficulty on Spatial Audio Target Acquisition in the Frontal Horizontal Plane

8.1 Introduction

In this Chapter, an attempt is made to design an evaluation methodology that can create design specifications for deictic spatial audio systems. Such a methodology should be able to specify design parameters that will result in usable interaction with a spatial audio display. In addition, this methodology should uncover context dependent effects such as mobility effects and also quantify the benefit induced by the acquisition task design.

In the previous chapters, the issues that have been found to affect interaction speed and accuracy were feedback design and Index of Difficulty. The evaluation technique should therefore quantify the effect of Index of Difficulty, result in estimating appropriate target sizes and provide insight on the success of potential feedback designs. Effects related to the context of use of the application should also be possible to determine using the evaluation technique.

With respect to Research Question 3 the study contributes by providing an evaluation design that is able to describe the acquisition task in many ways. With respect to Research Question 2 further effects of mobility and feedback as well as a detailed investigation onto the effects of target width and Index of Difficulty is undertaken.

8.2 Rationale

In Chapter 7, the selection task that was designed was found to be usable. Participants were able to perform the task successfully, in the presence of distracter sounds. However, a number of issues emerged that could not be explained solely based on the results of the previous experiment mainly due to the presence of the distracters. The results of the experiment however, imply that for particular values of target width, in addition to interaction accuracy interaction speed is affected even when Indices of Difficulty remain constant. Distance to target is expected to have a less significant role inside this target width range of values. Although target width values were sufficient for effective selection in the previous experiment, it is reasonable to assume that for lower values the effects that have been observed will be even more pronounced. For this reason, it is necessary to search the range of possible target width values and identify the area where performance is optimal. Optimality is defined in this sense as the target width and distance to target over target width range of values where interaction effectiveness is high and also interaction speed can be expressed in terms of Fitts' law as in visual target acquisition experiments.

Accomplishing this goal will enable a designer to decide on appropriate target width and distance to target values and expect that interaction performance in this range of values will be predictable based on the laws of visual target acquisition. To achieve this goal it is necessary to observe interaction in a wider range of target width values and ratios of distance-to-target to target width.

The success of the selection task presented in Chapter 7 was only verified with standing participants. Due to the fact that a promising application area for spatial audio displays is mobile human computer interaction, it is necessary to investigate whether the task can be successfully performed in truly mobile context. To examine this issue it is necessary to ask users to perform the task while walking in accordance with the literature in the field. In addition, it will be interesting to observe, the optimal target width and distance to target values for this particular case and how walking affects interaction speed and accuracy.

Furthermore, as has been found in Chapter 6, the success of the selection task is largely due to the feedback marked target areas. However, in this study target positions were confined in space. In this sense, it is necessary to investigate interaction in an egocentric display in more target locations and also in the absence of feedback to fully quantify the benefit induced by feedback marked audio elements and to decide appropriately on the potential of a non-feedback marked spatial audio display.

Finally, the evaluation technique that will be used can give answers to all these questions in a uniform and efficient way. Seeking a methodology that can reveal at once most of the findings from the previous experiments can lead to efficient evaluation of other types of pointing gestures, feedback designs or application contexts and result in design specifications that can be readily used for the development of a pointing based interactive system employing gestures and spatial audio.

Before proceeding with the experiment design, a review on studies that have addressed mobility on an experimental context is provided.

8.3 Mobility

Although a large number of studies exist on interaction with devices on the move, few look in a quantifiable way on the effect of mobility on interaction. Most seek ways to develop applications for people on the move and also identify application areas where such an attempt will be successful.

Given the fact that spatial audio displays provide an eyes-free way to interact, they are considered well suited for mobile HCI. Designing for mobile users, however, is considered a challenging task. The two major challenges are the limited display area that is available and the effect of mobility on the control actions of the user. As has already been mentioned, the design approach used in the thesis is to use spatial audio for display presentation and deictic gestures for control. Both of these choices help reduce the load on user's visual attention and are therefore suitable for mobile interaction. Gestures are a very convenient way of communicating in mobile situations. Most people can perform gestures whilst moving. For example it is very easy to point to something or to raise a hand to greet someone while walking.

Empirical evaluation by Pirhonen *et al.* [93] showed gestures to be a more convenient way of interacting with a system compared to common stylus based interaction techniques that require visual attention and inhibit our movement. A very common result in usability studies of systems supporting control based on visual feedback is that users have to interrupt their movement in order to interact with their computers.

Brewster *et al.* [19] found that an egocentric spatial audio display with directional head nodding gestures for selection could be used successfully while mobile, however no investigation on target size and Fitts' law was performed. Users were able to interact with the system whilst walking at an average of 69% of their normal walking speed. Empirical evidence therefore exists that gesture controlled spatial audio displays are usable in mobile contexts. In this experiment, the focus is however on pointing using the hand. It will be investigated how walking speed will be affected by target size as well as how mobility will affect selection times and selection accuracy. In addition, a comparison will be done to the standing case so that differences in performance can be estimated. Such a comparison had not been done in the Brewster *et al.* These results will help in understanding how display design has to be altered to compensate for the effects of mobility.

8.4 Experiment Design & Hypotheses

An experiment was designed to test the experimental hypotheses. Participants performed the task in Figure 31 according to the design presented in Table 16. Of the independent variables in Table 16, mobility, target width and A/W were tested within subjects, whereas feedback was tested between subjects. Participants were split in two groups, one performing the experiment with on target feedback and the other without. Both groups performed the experimental task standing and walking in a counter-balanced order.

Walking speed was calculated by dividing the number of laps participants performed with the total time this took. Preferred walking speed in this particular walking route was measured by asking participants to walk five laps without performing the audio selection task.

Dependent variables were time to select, selection success ratios, perceived workload (using NASA TLX workload sheets including an annoyance factor), percentage preferred walking speed and steps to select the target. Percentage Preferred Walking Speed is the percentage of participants' preferred walking speed they maintain when performing a specific task while walking.

DV	IV	Levels
Time Accuracy Walking Speed Steps Workload	Feedback	Yes/No
	Target Width	See Table 17
	Distance to target	See Table 17
	Mobility	Standing/Mobile

Table 16. Dependent and independent variables used in the experiment and their levels.

The research questions are concerned with the effect of feedback, mobility and Index of Difficulty on deictic spatial egocentric audio target acquisition in the horizontal plane. The hypotheses are:

1 Feedback in the form of audio on-target confirmation will affect accuracy of selections and interaction speed

2 Mobility will affect selection time and selection accuracy

3 Index of difficulty will affect selection time

8.4.1 Target Width and Distance to Target Manipulation

Due to the importance of target width on selection accuracy and efficiency, this variable is manipulated at six levels: 10°, 15°, 20°, 25°, 30° and 35°. Based on the results of Chapter 7 it would be expected that performance will be affected both in terms of both interaction effectiveness and interaction efficiency at these target width values. With respect to effectiveness it would be expected that the result will constitute part of the associated psychometric function with selection success ratios increasing as a function of target width until a ceiling effect is reached. With respect to efficiency it is expected that the difficulty incurred at the target width area of 10° will result in longer times to select the target even if Indices of Difficulty remain the same in the target width values. The magnitude of this effect is expected to diminish for larger target widths. The exact point after which the effect will become negligible will be determined by the results of the experiment.

A number of distances to target (A) are associated with each target width value. As can be seen in Table 17, variables are selected so that A/W ratios remain constant for most of the cases. Fitts' law would predict no significant differences in time measurements between target width values. However, based on the results of the previous experiment it is expected that differences will arise as a result of the difficulty that is associated with small target widths. The list of A/W ratios examined is small. This is mostly done to restrict the number of trials each participant has to perform but also due to the fact that for the particular task it is hard to obtain high ID values. This is due to the restricted display area in our study (the area in front of the user) and the relatively large target sizes that have to be used in order to provide usable pointing. For this reason, for target widths of 10° and 35°, it was decided that tests should be

performed for additional values of ID, namely 3 and 2.48. In the former case, this was done to gain some insight on what is happening at higher IDs and in the latter because using an ID of 2 would result in a distance that would lie to the back of the user whereas our test area was in front. It should be mentioned here that the concept of target width only affects user performance when participants get feedback. For the group where no feedback was given, ID is not expected to affect the results. The relevant variable in this case is distance to target.

To examine the hypotheses outlined and the applicability of Fitts' law (at least for the particular indices of difficulty) on spatial audio target acquisition it was decided to examine the effect of target width and distance to target at the levels presented in Table 17.

W	A	ID	A/W
10°	10°	1	1
	30°	2	3
	70°	3	7
15°	15°	1	1
	27°	1.5	1.83
	45°	2	3
20°	20°	1	1
	36°	1.5	1.83
	60°	2	3
25°	25°	1	1
	46°	1.5	1.83
	75°	2	3
30°	20°	1	1
	55°	1.5	1.83
	90°	2	3
35°	35°	1	1
	64°	1.5	1.83
	87°	1.8	2.48

Table 17. Target Widths (W), Distances to Target (A), and Indices of Difficulty (IDs) and A/W Ratios used in the experiment.

8.5 Experimental Task

The task participants had to perform was to select the spatial audio target, in the associated combination of the experimental conditions. Distance to target and target width for each trial was selected

randomly out of one of the rows presented in Table 17. Participants repeatedly performed the task until 8 observations from each row of Table 17 were obtained. Overall each participant performed 108 selections standing and 108 walking. Depending on the experimental condition combination they were either standing or mobile and received or did not receive feedback. To select the target sound participants had to move their hand in the direction of the target sound and perform a downwards gesture with their wrists to indicate selection. Audio feedback was provided to indicate that the selection gesture was recognized by the system. Participants performed the selections one after the other in a serial fashion, the starting point for each gesture was the terminating point of the previous one. When mobile, participants had to walk in figure of eight laps around three traffic cones that had been placed in the lab. The cones were 1.2 meters apart, providing a rather challenging route. This was done to provide a realistic scenario forcing participants to pay attention to their movement.

Figure 39. A participant selecting a virtual 3D sound source while walking and wearing the experimental apparatus.

8.6 Stimuli & Apparatus

The stimulus was half a second of white noise repeating itself after half a second of silence, played at the height of the user's nose at a distance of 5 meters. Feedback was provided to the appropriate group by the sound of people cheering when participants were inside the target width that was assigned to each

trial. The feedback sound was played from the same direction as the target when participants entered the target area.

To create a truly mobile test bed a small laptop was used that was placed in a rucksack that participants wore on their backs and ran the program used to control the hardware and the stimuli. Each participant was equipped with two MT-9B (www.xsens.com) orientation trackers. One of them was placed in a small belt-mounted case that was placed in the middle of each user's waist. This tracker was used to record user movement and calculate the number of steps taken. The other was held in the right palm of each participant. The difference in orientation readings between the two trackers was used to infer pointing direction of the hand relative to the body. In this way, an estimate of pointing direction was available while participants could freely move in space. HRTF filtering was done on the laptop using the DieselPower Implementation 3D Audio API (www.am3d.com). The API provides an HRTF filtering implementation that uses generic (non-individualized HRTF functions). Localization using non-individualized HRTFs is worse compared to localization using individualized HRTF functions, however the effect in sound localization on the frontal horizontal plane is non-significant. Sennheiser HD 433 headphones were used to present the sounds.

8.7 Procedure & Participants

Participants were assigned to one of the two groups and were briefed on the experimental task. They were asked whether they were facing any hearing deficiencies and if they did they were excluded from the experiment. Participants were instructed that once the experiment started they would experience a sound which would be perceived as playing from a fixed position somewhere between their left and their right and always in front.

They were told that their task would be to move their hand to the position of the target as fast as possible and perform a downward hand gesture to indicate the sound position. When participants were part of the feedback group they were told to make sure they heard the feedback sound before proceeding with their selection.

In total, 24 participants were tested 6 females and 18 males with an age span of 18-42 years (mean 27). All were students from the University of Glasgow and were paid £5 for their participation. Minimal training with respect to the use of the equipment was given to participants prior to performing to the task.

8.8 Results

The analysis presented here is based on the raw data that are available in the associated section of Appendix 1. Due to target width having no affect on interaction in the no-feedback case, given that it was not implicitly perceptible an overall mobility (2) x feedback (2) analysis of variance was performed first. Mobility was found to have a significant main effect on time to select a target ($F(1,2590)=141.24$, $p<0.001$). Feedback was also found to have a significant main effect on selection time

(F(1,2590)=1134.909, p<0.001). Mobility was found to have a significant main effect on the accuracy of selections as measured by the absolute deviation from target, F(1,2590) = 616.054, p<0.001. Feedback was found to have a significant main effect on the accuracy of selections, F(1,2590)=1148.477, p<0.001.

Participants were slower and less accurate when mobile. They were slower but more accurate when they received feedback. Given the main effects observed, the thesis proceeds with two separate analyses for participants performing the task with and without feedback. Results are illustrated in Figure 40. It is interesting to observe that when considering mean accuracy and speed values the improvement on accuracy is much higher that the degradation in speed. Accuracy was four times worse when no feedback was given, while the decrement in interaction speed induced by feedback was only about 40% for standing participants. For mobile participants accuracy was also improved by a factor of about 3 while interaction speed was increased twice. This shows that the provision of feedback benefits interaction performance both in standing and mobile situations.

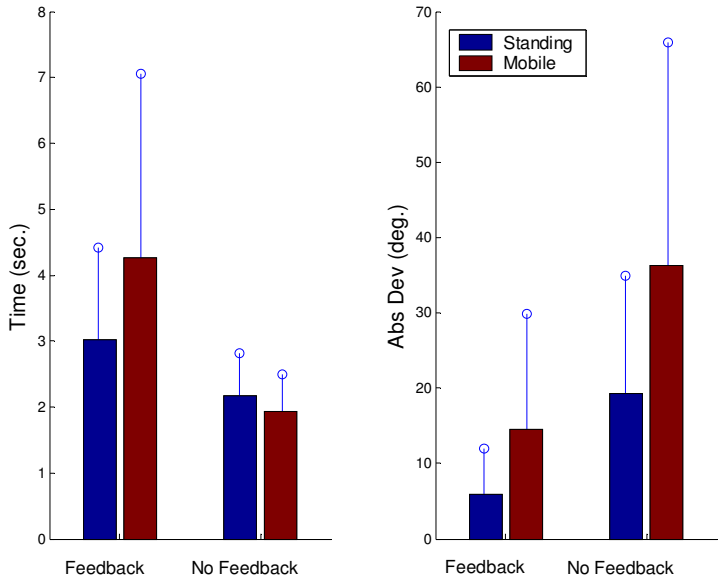


Figure 40. Means and standard deviations of the time and accuracy scores for standing and mobile participants who did and did not receive feedback

8.8.1 Performance with on-target feedback

8.8.1.1 Time Analysis

Figure 41 presents mean selection times for participants that received feedback as a function of mobility, target width and Index of Difficulty.

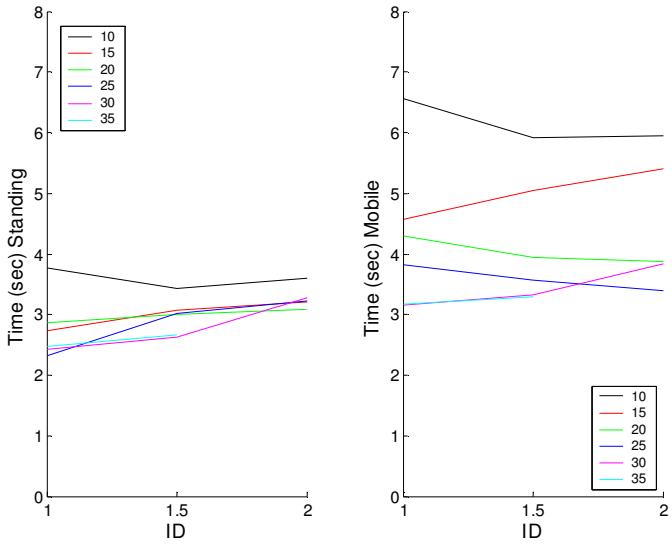


Figure 41. Mean time scores for participants who received on-target feedback as a function of ID and target width.

An overall mobility (2) x width (6) x A/W (3) analysis of variance was performed on the time scores of the participants that received feedback. Results are presented in Table 18. The results confirm a significant main effect of mobility, width and A/W ratio. Interaction between mobility and target width as well as between width and A/W ratio was also significant. For standing participants *post-hoc* t-tests using the Bonferroni confidence interval adjustment reveal time scores for target width of 10° to differ significantly from all the rest, with no other differences found. Time scores for all A/W ratios were found to differ significantly for standing participants.

	T	S
M	F(1,143)=34,619,p<0.001	F(1,11)=122.335,p<0.001
W	F(5,715)=38.071,p<0.001	F(5,55)=23.861,p<0.001
A/W	F(2,286)=11.421,p<0.01	F(2,22)=4.690,p<0.02
MxW	F(5,715)=17.230,p<0.001	F(5,55)=23.861,p<0.01
MxA/W	F(2,286)=5.819,p<0.01	N/S
WxA/W	N/S	N/S

Table 18. ANOVA results for participants that received feedback (T is time and S (%) success ratio). M stands for mobility, W for target width and A/W for the ratio of distance to target to target width.

For mobile participants A/W ratios did not have a significant main effect on selection times. Pair-wise comparisons showed selection time for widths of 10°, 15° and 20° to differ significantly from all of the other target widths and between themselves. There was no difference in selection times between target widths of 25° and 30° which, however, differed from all the rest; no difference between 30° and 35° and 25° and no difference between 35° and 30° target widths were found. Figure 42 shows time scores for all tested target widths averaged over A/W ratios.

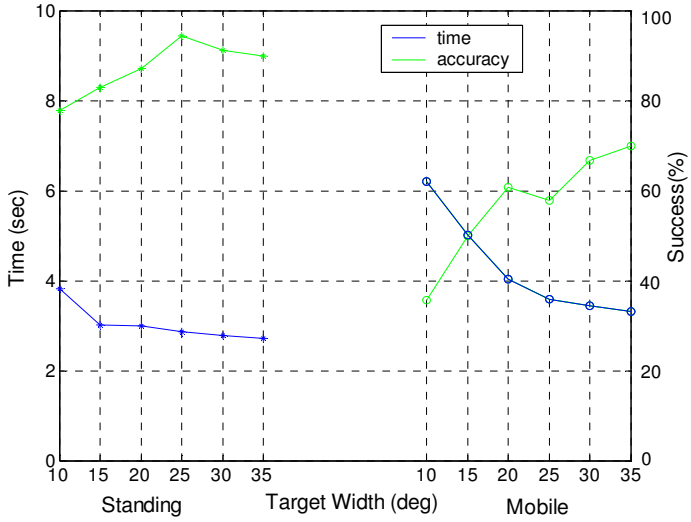


Figure 42. Time & Success scores as a function of target width for standing and mobile users who received on-target feedback averaged over A/W ratios.

8.8.1.2 Success Ratio Analysis

An overall mobility (2) x width (6) x A/W (3) analysis of variance was performed for success scores for participants who received feedback. Selection success scores refers the percentage of trials participants selected within the feedback marked area. Results are presented in Table 18. The results confirm a significant main effect of mobility, width and A/W ratios. Success ratios increased with target width and decreased when participants when mobile. Interaction between mobility and target width was found to have a significant effect on success scores. *Post hoc* analysis comparing mobile and standing participants showed A/W ratios to have a significant main effect on success ratios for mobile but not for standing users. Figure 42 shows how success ratios varied for the aforementioned cases, averaged over A/W values.

8.8.1.3 Steps Analysis

A width (6) x A/W (3) analysis of variance was performed on the number of steps taken per selection for (mobile) participants who got feedback. The results showed width to have a significant main

effect F (5,355) = 23.836, p<0.001. No effect of A/W ratio was found. Figure 43 presents the results grouped over A/W ratios. Grand mean was 5.6 steps.

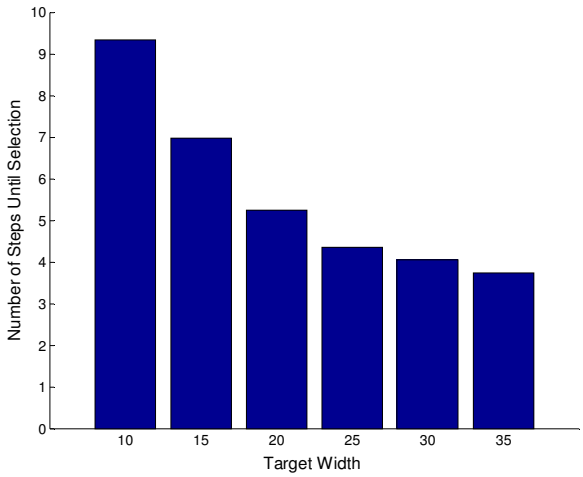


Figure 43. Mean steps until selection as a function of target width.

8.8.2 Performance without feedback

Without feedback target width was not evident to participants and therefore is not a relevant variable. The factors affecting performance in this case are distance to target and mobility. A within-subjects analysis of variance (mobility (2) x distance (15) (see Table 2) on time scores reveals a significant main effect of both mobility and distance (F(1,71) = 31.040, p<0.001 and F(17,1207) = 2.497, p<0.005).

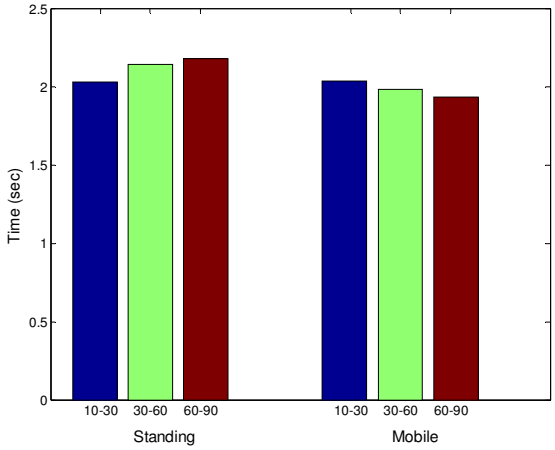


Figure 44. Mean time scores for participants that did not receive on-target feedback as a function of distance to target groups and mobility.

Post hoc, pair-wise Bonferroni confidence interval adjusted t-tests showed interaction to be faster for mobile participants compared to standing ones (see Figure 44). Means of the three distance groups that result from averaging time scores every six distances (with respect to Table 2) are presented. A similar analysis on standard deviation from target revealed a significant main effect of mobility with participants being less accurate when mobile, as shown by Bonferroni *post hoc* t-tests. No effect of distance was found in the no feedback case. Figure 45 shows the selection accuracy for each target width when standing and mobile. Participants were significantly less accurate when mobile. It can be observed that very large target widths would be required for effective interaction under these conditions. With respect to steps taken to select overall distance to target produced a significant main effect ($F(17, 1207) = 2.2111$, $p=0.003$). On average, participants performed about two (1.74) steps before proceeding to each selection.

8.9 Walking Speed Analysis

An analysis was performed to test the effect of feedback on percentage preferred walking speed for mobile participants. No significant difference was found between participants using feedback and no feedback. Both, however, were significantly less than participants' normal walking speed.

Means were 71% for the no feedback case and 76% for the feedback case, a finding consistent with [3].

8.10 Workload Analysis

A mobility (2) x feedback (2) analysis of variance showed mobility to have a significant main effect on overall perceived workload ($F(1,22) = 37.498$, $p<0.001$), with mobility increasing perceived workload by about 30%, (Mean Standing Workload = 3.65, Mean Mobile Workload = 4.8, max 10). Feedback did not have a significant main effect on overall perceived workload. A more detailed analysis showed mobility to have a significant main effect on mental demand ($F(1,23) = 13.089$, $p<0.001$), physical demand ($F(1,23) = 5.741$, $p<0.03$), effort expanded ($F(1,23) = 12.032$, $p<0.005$), performance level achieved ($F(1,23) = 9.304$, $p<0.01$) and frustration experienced ($F(1,23) = 13.301$, $p<0.001$). Mobility did not affect time pressure and annoyance experienced.

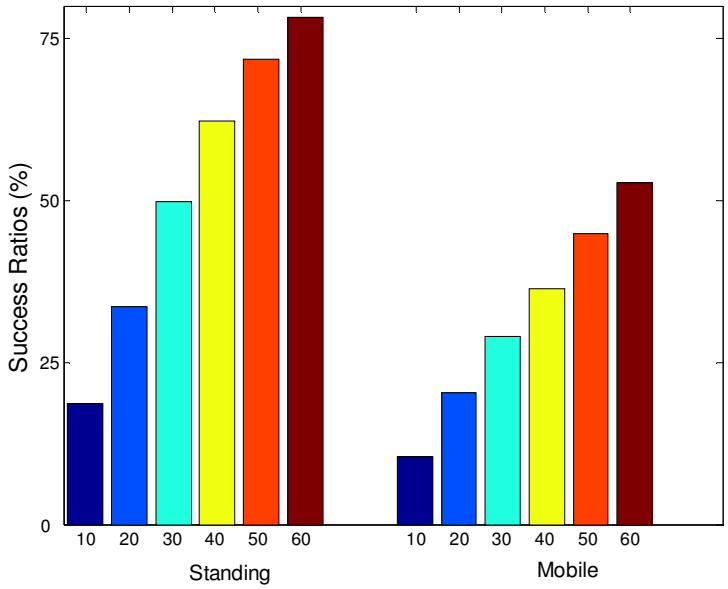


Figure 45. Success ratios as a function of target width for standing and mobile users (no on-target feedback)

8.11 Can Fitts' law be used to describe 3D audio target acquisition?

Is it possible to model spatial audio target acquisition in terms of Fitts' law? This analysis is performed only for participants that received feedback. Linear regression on time measurements was performed for the three models in Equations 1, 2 and 6 for W values in the usable range that is over 20° for standing participants and over 30° for mobile ones. These values were chosen because no significant difference was observed in time ratings as can be seen in the results section. Regression results are presented in Table 19 and Figure 46.

For standing participants, it can be observed that both logarithmic models correlated significantly, the model of Equation 2 correlating significantly better as the z statistic, using Fisher's Z transformation of the correlation coefficients, revealed ($z = 2.58$, $p<0.05$). The linear model, although providing high correlation, did not correlate significantly ($p = 0.058$). Given the significant correlation values the Index of Performance can be calculated for standing participants. This was approximately 1.6 which is about half what has been found in the literature for participants interacting with a visual display using a trackball [13]. For mobile users this calculation was not made since the models did not correlate significantly. Throughput values were also calculated for standing participants giving values of 0.4, 0.53 and 0.63.

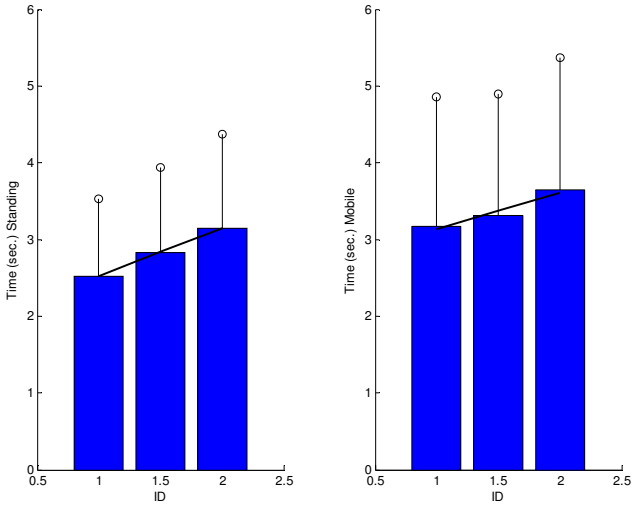


Figure 46. Linear Regression lines, mean time scores and standard deviations as a function of Index of Difficulty

This is about 3.5 times less than that measured for interaction with virtual pointers in visual displays. Throughput was calculated using A/W values and not the effective target width formulation because effective target widths were unreasonably high and dependent on target size. For example, for target width of 25°, effective target width was 36° and success ratio in the order of 95%, 36° therefore does not provide a better estimate of target width. Throughput increased for increased distance to targets. Indeed, the time measurements reveal that, at the lowest IDs, participants became confused since the feedback area was too close to their initial position. Time scores in throughput calculations include reaction time and selection time (therefore throughput values might be underestimated).

	R^2_{Lg}	R^2_L	R^2_F	IP
S	0.999*	0.991	0.995*	1.57
M	0.86	0.92	0.82	X

Table 19. Goodness of fit comparisons between the linear and logarithmic models. * denotes that correlation was significant at the p<0.05 level. R^2_{Lg}, R^2_L, R^2_F stand for the R^2 statistic for McKenzie, Linear and Fitts' models. IP is the index of performance. S stands for standing and M for mobile participants.

8.12 Discussion

The results of this study verify to a great extent the experimental hypotheses. Mobility was found to cause slower and less accurate interaction. Feedback was found to decrease interaction speed but increase interaction accuracy and interaction was negatively affected by increased Index of Difficulty. The success ratios in the no-feedback condition show that performance in deictic interaction with an egocentric display without feedback is very poor. It can be observed that such a display can hardly accommodate more than three targets in the area in front of the user. In mobile situations, the maximum number of targets would be two. It has to be said however, that target position was varying randomly in this experiment, in this sense the effects of learning are not taken into account in the results. It might be the case that when interacting with a familiar display one more target might be feasible. Surprisingly, when mobile, participants were faster in their selections when on-target feedback was not given. Participants commented that when mobile they were not able to pay much attention to the target position and they mostly selected on a left right basis. They said that when standing they were able to pay more attention to the target sound and infer more on its position. Consequently, they adopted a more careful strategy for their selections that resulted in increased selection times. This finding is therefore attributed to the negative attitude participants formed towards the system when mobile.

On-target feedback was found to improve performance significantly for standing users and, based on the results of this study, is considered necessary to enable usable gesture interaction with a spatial audio display. Feedback design is, however, critical. For standing users, it was found that time to select is not affected by target width for widths of 15° and more. The fact that target time is affected by A/W scaling is not novel and was verified in studies in the visual domain [40, 52]. Based on out results this effect is minimized for target widths over 15°. However, target widths between 20° to 25° were found to be necessary to provide successful selection accuracies of 90% and more. On these grounds, target widths of 20° or more are recommended for standing users. Consequently, such a display could accommodate up to 8 elements in front of a user. On-target feedback was found to be moderately successful in the case of mobile users. It did not affect their walking speed compared to the case of no feedback, nor did it increase the perceived workload. However, as shown in Figure 42 success rates of more than 75% were not possible to achieve for the target widths that were tested in this study. Although time to select when mobile with a target width of 35° was relatively close to the one observed for standing users, it appears that wider targets than this are necessary to increase success rates. Observing Figure 42 it can be deduced that the performance curve has not saturated and increasing target width further will benefit interaction both from an accuracy and speed point of view. Therefore, mobility was found to influence performance in the feedback case in a negative way. Even with large targets there is a 20% increase in selection times and an approximately 20% decrease in selection accuracy. Increasing target width is a solution as found in this study, however this approach is not optimal because it reduces the number of elements the display

can accommodate. Alternative ways to overcome the variability introduced in mobile contexts have to be examined to reduce the negative effect that was found in this study. Appropriate filtering of the movement signals is a promising solution for overcoming mobility problems and provide an experience that will resemble the standing case more closely.

The modalities chosen for display presentation (3D audio) and control (gestures) have been proven in the study to provide a mobile way to interact with the system. Users did not have to stop at all when performing the task, neither with nor without feedback. As can be seen in the number of steps to select the target analysis at appropriate target widths users were able to select a target approximately every three steps, a promising finding given the relatively complicated walking route they had to follow. In addition, the resulting percentage preferred walking speed is close to the preferred one, with a mean of about 73%. Given the random target positions in this experiment, even higher figures can be anticipated for users interacting with a familiar system.

According to the results presented, it appears that it is possible to view spatial audio target acquisition in terms of Fitts' law. This is particularly encouraging since it shows that this type of interaction is an efficient one, comparable to interaction with visual displays. This issue is of particular importance since it enables cross-modal comparisons in the context of tasks under examination. In addition, it provides a predictive tool for performance in gesture interaction with spatial audio. Even in the case of mobile participants, high, although not significant, correlations were observed. The complex walking route participants had to follow definitely contributed to this finding. In another context involving a simpler walking route it might prove feasible to calculate the difference in the index of performance between standing and walking conditions. Further design is necessary in order to create pointing tasks that can be described by Fitts law in mobile contexts. In conclusion, this study revealed a number of major factors that affect performance in deictic interaction with a spatial audio display. The selection task that was proposed and examined was found to be usable for standing users and to allow mobile users to interact with a walking speed close to their normal one, while at the same degrading selection success rates and time at the level of 20%. Given the elementary design that was employed in the experiment the results enable us to be optimistic on the future of gesture interaction with spatial audio in mobile contexts.

8.13 Conclusions and Guidelines

All experimental hypotheses were verified. Feedback affected time and accuracy of selections, mobility affected interaction speed and accuracy and index of difficulty affected time to select a target.

The results of an empirical study that showed mobility, feedback and Index of Difficulty to have a significant effect on spatial audio target acquisition were presented. The acquisition task design was found to result in usable deictic interaction with no effect on walking speed or workload compared to interaction without feedback. Spatial audio target acquisition was found to be sufficiently described in

terms of Fitts law, when proper target width choices were made. A detailed investigation on the effect of target size on time and accuracy of selections was presented. Participants were able to walk at 73% percent of their normal walking speed, mobility degrading performance by 20%. Deictic interaction with 3D audio displays is shown to be a feasible solution for future human computer interaction designs.

With respect to Research Question 1 this chapter contributes by showing that the acquisition design that was proposed does not affect perceived workload and can be used by mobile users. With respect to Research Question 2 novel findings were presented on the effect of the Index of Difficulty, the effects of mobility and feedback deprivation. Finally, with respect to Research Question 3 the evaluation technique proposed was justified since the acquisition task complied with Fitts' law. An Index of Performance of approximately 2 can be expected for standing participants which is about half the one observed for visual interaction using virtual pointers.

Based on the results of this study the following recommendations to designers can be made:

- Both in standing and in mobile situations, providing feedback marked audio elements will improve interaction and will not affect perceived workload or walking speed.
- Designers should expect an approximately 20% deterioration of performance when people using their system are mobile. They could compensate by accordingly increasing target widths in the system.

9 Conclusions

9.1 Introduction

In this chapter, a summary of the results of the thesis and its contributions will be presented. In addition, the limitations of the thesis are discussed and proposals are made on how they could be overcome in future work.

9.2 Summary of the Thesis

The thesis began by reviewing the literature in the field of spatial audio display design and identified that the question of how direct manipulation principles could be applied in deictic spatial audio display design has not been examined. In Chapter 3, it was found that the following properties must be supported for auditory direct manipulation to be usable:

- Adequate differentiation of display elements in a spatial sense
- Support for control based on aimed movements
- Creation of successful semiotic mappings to communicate a valid conceptual model
- Support for closed loop control based on audio feedback to help users to evaluate the results of their actions and validate their conceptual models

According to the literature only the third has been partly researched and evaluated in the field of audio display design. This, however, has mostly been done in the context of creating auditory semiotic information and the validation of whether and to what extent a valid conceptual model can be supported by spatial audio semiotic information is still a field under investigation. The thesis did not address this issue but rather focused on the first two points: whether spatial audio can support adequate differentiation of display elements in a spatial sense and whether it can support control based on aimed movements.

To obtain further insight on the issue of aimed movements, the thesis examined visual target acquisition with two goals. The first was to identify what are the perceptual and motor skills that support aimed movements and the second to identify an evaluation methodology that can be applied on a spatial audio acquisition task. With respect to the first goal according to the review presented in Chapter 4, aimed movements to visual targets are performed by means of an initial ballistic movement followed by corrective submovements. The literature indicates that visual and proprioceptive feedback throughout the aimed movement contributes to the successful accomplishment of this task.

With respect to the second goal it was found that vision supported aimed movements in ‘normal’ ranges of difficulty are characterized by the logarithmic model proposed by Fitts and revised by

MacKenzie. Although linear as well as power models exist, these do not apply in the case of the spatially constrained tasks that are typical in Human Computer Interaction. The thesis therefore opted to obtain time measurements for spatial audio target acquisition task and to verify whether they would conform to Fitts' law. In relation, however, to the definition of usability, apart from examining the aforementioned issue, it is of importance to examine the effectiveness of the acquisition technique in terms of selection success rates as well as user satisfaction. Due to the early stage of development in the field the thesis is investigating the latter issue is handled by estimating mental workload using NASA TLX forms [54].

The thesis proceeded by examining available psychoacoustical literature, in Chapter 2. According to the findings, it was concluded that pointing to a spatial audio target lacks the attributes that make pointing to visual targets so successful. Contemporary spatial audio systems use non-individualized HRTFs and are prone to increased localization error as well as front-back and up-down confusions. It appears that the attribute which can be judged most consistently, within known values of localization error, for a wide population without training is azimuth. In addition, a major issue that was found is that the border separating a spatial audio sound from its background can be ambiguous.

These findings result in certain limitations for a designer. The first is that display area should be confined to the frontal horizontal plane. Such a decision can successfully overcome problems related to confusions and in addition result in consistent judgements of sound direction for a wide class of users. However, even in this case there is no way to indicate to the user which area is associated with each audio display element by the system.

As the review indicates, the research questions that are raised by the thesis are novel and have not been examined in the literature. The research questions are repeated here.

RQ 1	How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?
RQ 2	What are the factors that affect deictic spatial audio target acquisition?
RQ 3	How can we evaluate the usability of deictic spatial audio target acquisition?

With the goal of designing of a spatial audio task that overcomes the limitations of directional hearing, the thesis proceeded with experimental investigations in this field. Four experiments were presented, the first two concerned with obtaining the necessary information with respect to the design of a spatial audio target acquisition task and the last two with the evaluation of this task.

The first experiment addressed the problem of estimating target size in spatial audio target acquisition as a function of the gesture used. This was a feasibility study that was helpful in identifying and verifying the problems that would be expected in deictic interaction with a spatial audio display. The study used an exocentric display, where participants did and did not receive feedback and used three different gestures to browse and select a single target sound in the soundscape. The results showed that

interaction was unfeasible for a large number of participants that did not receive feedback. Interaction with feedback proved to be feasible and estimates of target widths that result in selection effectiveness in the order of 70.7% were obtained. It was found that the resolution and movement support offered by the use of a stylus on a touch tablet was the most beneficial to the effectiveness of target acquisition. It was followed by selections using the hand and selections using the head of the user. However, mostly due to the exocentric nature of the display, the success of the tablet solution was compromised by the fact that participants did not find it easy nor comfortable to perform compared to hand and head acquisition. Due to the fact that more research was found to be necessary to design a spatial audio target acquisition task, it was decided to proceed in the next experiments using the hand gesture, because it could performed naturally and resulted in high effectiveness.

The aim of the second experiment was to examine different feedback types as means of enhancing the usability of deictic interaction with a spatial audio display and identify the effect of distracting sounds on interaction. The results of the experiment showed a speed accuracy trade off when comparing egocentric and exocentric interfaces. In the former case, interaction is fast but not accurate, whereas in the latter it is accurate but slow. Imposing a loudness cue did not help in terms of interaction efficiency and effectiveness. Audio marking the target area, however, was found, to compensate for accuracy errors in egocentric and interaction speed in exocentric interfaces. The effect of distracting sounds only appeared in the case where the feedback cue was continuously mapped to the pointing direction. In this case interaction speed was affected in a negative way, however, no effect on interaction accuracy was observed. Interaction in an egocentric display with feedback marked target areas is therefore a promising solution where interaction performance is not affected by distracting sounds.

With the goal of designing a spatial audio target acquisition task that could be applied in a wide variety of contexts, it was decided to examine deictic interaction in an egocentric display where target areas are feedback marked. This task was evaluated in Chapter 7, in a challenging experimental design where distracter sounds that had the same timbre as the target sound were used. It was found to be usable and in addition possible to perform using alternate presentation modes that alleviate the problem of user isolation for the real world auditory environment. These modes were bone conductance headphones and monaural presentation. Interaction accuracy was not affected by presentation mode and interaction speed was reduced only in the case of monaural presentation due to the restricted localization cues. In addition, the experiment provided promising results in the context of modelling spatial audio target acquisition using the models of visual target acquisition. Under headphone presentation high correlation with a logarithmic model was observed, however the results were not conclusive due to the presence of distracter sounds. According to the results of the experiment, target width was found to be more important than distance to target in the context of the specific acquisition task.

The thesis concludes in Chapter 8 with an experimental design that can be used for a variety of feedback types and gestures. In the last experiment, the target acquisition task is examined in detail and

the effects of target width that were found in Chapter 7 are quantified. The acquisition task that was designed is found to be usable when mobile. In the absence of feedback, although interaction was faster, the inadequate support for aimed movements leads to high accuracy errors that restrain the scalability and applicability of this option. The importance of feedback is stressed by the findings of this study, and in addition, feedback is found not to affect walking speed nor perceived workload. A reduction in performance in the order of 20% was observed when people were mobile. Finally, the acquisition task was found to comply with Fitts' law, a fact that provides additional support for its usability. An index of performance in the region of 2 should be expected for standing people interacting with spatial audio targets, which is about half the one observed for visual interactions using virtual pointers.

9.3 Thesis Contributions

The major findings of the thesis are the design of a spatial audio target acquisition task and the formulation of a methodology for its evaluation. With respect to the Research Questions that were set in the beginning of the thesis, the following answers can be given at this point.

RQ 1: How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?

In order to support spatial audio target acquisition, it is necessary to overcome the problem of localization error and the problem of confusions with respect to sound direction. The latter can be overcome by restricting the display area. In order to find out how the former problem can be resolved it is necessary to identify what are the cues that constitute visual target acquisition successful. Visual target acquisition is supported by the detailed feedback provided by vision with respect to the target area. This information is used both to support the first ballistic submovement and also to support the corrective submovements that guaranty selection effectiveness. The more detailed the information given by the feedback cue the better the accuracy and speed with which users will select the spatial audio target. Therefore to overcome this problem it is necessary to 'mark' the target area. Feedback marked audio targets were the best solution to compensate for the problems of directional hearing. This finding was supported by the success of the acquisition task that was proposed in Chapter 7 and the negative effect on performance that was observed under feedback deprivation in Chapter 8.

RQ 2: What are the factors that affect deictic spatial audio target acquisition?

The experiments in the thesis identified a number of factors that affect spatial audio target acquisition. They showed that deictic spatial audio target acquisition is affected by the pointing gesture that is used, the type of feedback cues that are given to users with respect to target sound position, the size

of and the distance to the target area, the quality of the localization cues that are provided and the context of use in the case of mobile users.

It was found that performance quickly deteriorates in feedback deprivation situations. In an egocentric display is feedback is not provided with respect to target position, accuracy is worse by a factor of almost four while the benefit in speed is only in the order of 50%. The thesis also provided quantitative measurements on the effect of target width size on the speed and the accuracy of the acquisition of feedback marked spatial audio targets. With reference to the experiments and the technology used in the thesis it was found that spatial audio target acquisition in an egocentric display is invariant to scaling A/W ratios for target widths over 15° however a target width of 20° is necessary to result in accuracy scores higher than 90%. It should be noted here that the effect of the equipment and the gesture used for pointing is also dramatic, given that when using a touch tablet the required target size was almost half compared to using a physical pointing gesture using the hand and or the head. Mobility results in degradation of performance in the order of 20%. Finally, it was found that the task of selecting a spatial audio sound when walking results in a walking speed that is about 73% the one people have under normal conditions.

In addition, it was found that the quality of the localization cues is crucial when the spatial audio element sizes are feedback marked. In particular, the feedback cue ensures that interaction will be effective. With respect to the efficiency of interaction it was found that a stereo like cue similar to the one induced by bone conductance headphones can efficiently guide selections for sounds in the frontal horizontal plane, which is explained by the fact that azimuth localization in this area is dominated by interaural cues. Monaural presentation was found to degrade interaction speed by a factor of 2.

Finally, the thesis found that there is a duality between interaction in egocentric vs. exocentric interfaces. When no other feedback is given to participants and in the context of our experimental task, participants were three times less accurate in the egocentric display, being two times faster at the same time.

RQ 3: How can we evaluate the usability of deictic spatial audio target acquisition?

Based on the results of the thesis, it can be concluded that the usability of a particular spatial audio target acquisition technique can be evaluated in a manner similar to the one used in visual target acquisition studies. It is necessary to examine interaction speed, interaction accuracy and user satisfaction. Furthermore, the notion of the Index of Performance was found to be applicable in the case of feedback marked spatial audio targets and in this sense can be used to compare between future target acquisition techniques. The thesis also found that Fitts' law can model the acquisition of feedback marked areas. The thesis found that the proposed acquisition task is about half as efficient, compared to visual

interaction with virtual pointers and these findings indicate that it is possible to perform cross modal comparisons based on objective quantities.

A number of guidelines were derived from the experiments performed in the thesis. These are brought together here, as follows:

1 Both in standing and in mobile situations, providing feedback marked spatial audio elements will improve interaction performance and will not affect perceived workload or walking speed. This is true for both egocentric and exocentric interfaces.
2 Target size in an egocentric display controlled by pointing using the hand of a standing user should be more than 20°.
3 When keeping contact with the real world audio environment is important, designers should consider the possibility of monaural or bone conductance presentation of the display. In the latter case, interaction speed and accuracy should remain the same. In the former, there is a decrease on interaction speed by a factor of 2.
4 Natural options of browsing a real time updated virtual auditory environment, such as head movements, are preferred by users.
5 Increasing the number of display elements will not significantly influence interaction performance in familiar egocentric displays as long as no intelligibility problems appear. In exocentric displays, however, because of the higher cognitive load imposed by the real time updated localization cues, increasing the number of display elements affects interaction speed in a negative way.
6 When selecting a feedback marked spatial audio element using a stylus operated touch tablet 9° around the target should result in 70.7% selection success rate. This number becomes 16° and 18.5° for selecting using the hand and head of the user respectively. In the hand case, this number refers to selections in front of the user.
7 A designer should not rely on a loudness based cue to deliver directional information. Such a cue is useful in improving intelligibility rates however, its success as a directional cue is limited.
8 Designers should expect an approximately 20% deterioration of performance when people using their system are mobile. They could compensate by accordingly increasing target widths in the system.
9 When possible, designers should allow for display size to grow, increasing target width since its effect is beneficiary from an accuracy point of view, and its effect is more important than that of distance.

In conclusion, the thesis contributes to the field of spatial audio display design by disambiguating the task of pointing to an audio target. Prior to the thesis, this aspect of interaction with a spatial audio

display had not been examined. The thesis identified novel effects of gesture, feedback, mobility and index of difficulty on spatial audio target acquisition. In addition, the thesis provided estimates for certain parameters of display design, especially with respect to target width. Furthermore, the thesis contributed with the design of a spatial audio acquisition task that was found to be usable both for standing and for mobile users and can be performed using alternative reproduction techniques that alleviate the problem of user isolation from his/hers real audio environment. Finally, the thesis contributed with an evaluation technique that can be used to evaluate and obtain design information on a number of different designs and feedback types in a fast and uniform manner.

9.4 Thesis Limitations & Future Directions

In general the undertakings in the thesis did provide useful insight with respect to the research questions that were set. However, there are a number of limitations that need to be considered. It should be noted firstly that the thesis is limited in the context of spatial audio display design because no consideration has been made with respect to sound design for spatial audio displays, in the sense that it did not examine the extent that auditory semiotic mappings can support a valid conceptual model. This is due to the fact that such a study is too wide to be considered in the context of the thesis. Although the thesis was successful in providing a pointing task for spatial audio displays, in order for the field of spatial audio display design to develop it is necessary for further research to be performed with the goal of evaluating auditory semiotic information and display presentation issues. The discussion on the limitations of the thesis will proceed by firstly examining the limitations with respect to the answers that were given to the research questions.

RQ 1: How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?

The thesis found that the major problems with spatial audio target acquisition result from the fact that the auditory image that we create for a virtual ‘sounding’ object does not convey clearly information with respect to the exact position of the object or its size in the display. The thesis compensated for this problem by marking the object’s area with audio feedback. However, the thesis had to constrain the study in the frontal horizontal plane to avoid the problem of confusions, which restricts the display area to the front of the user. In this sense, the display area that could be used based on the findings of the thesis is limited and might not be enough for particular applications.

RQ 2: What are the factors that affect deictic spatial audio target acquisition?

The thesis successfully identified a number of factors that affect deictic spatial audio target acquisition, in particular with respect to the usability of the selection process. However, the thesis did not look into aspects related to sound design and the effects of simultaneous presentation of sounds. The problem of how to create a soundscape that can be used for everyday tasks is a rather challenging one since it has to satisfy aesthetical, intelligibility and usability aspects at the same time. A user is aware of an audio display all the time and there is no way for him to close his ears to elements that are not of interest to him. In this sense the design must be such, that it does not overload the user and at the same time supports active exploration by ensuring the 'audibility' of the display elements. The thesis does not address this issue directly and in this sense factors related to this aspect of design would be expected to influence performance in ways that have not been addressed in the text.

In addition the thesis did not investigate in detail the gestures that were used in the context of Chapter 5, although these were promising in terms of selection effectiveness. Further study using the experimental design developed in Chapter 8, could be performed to show how exactly these gestures could be applied in spatial audio display design.

Furthermore, although the designed selection task was found to comply with models of visual target acquisition, a detailed investigation was not performed for the feedback designs that were used in the experiment performed in Chapter 6. The reason is that at that point of thesis development it was necessary to quickly obtain information on the major differences between the different feedback types and not model arm movement in the presence of the evaluated feedback. The utilization of the notion of throughput in the context of that experiment is therefore arbitrary and awaiting further evidence to show whether the notion of the index of difficulty when considering the tasks in the particular experiment is an appropriate quantity. However, this metric was found useful in providing a uniform measure for the evaluation of the feedback designs under consideration. Future work could well investigate whether such a model of the index of difficulty is appropriate for interaction using the feedback cues evaluated in Chapter 6.

Finally, all the experiments in the thesis did not assess the effects of practice. This issue is quite interesting since performance is expected to improve and in this sense estimates of design parameters that would suit skilled users could be obtained.

RQ 3: How can we evaluate the usability of deictic spatial audio target acquisition?

The thesis found that the acquisition of feedback marked spatial audio targets can be evaluated in a similar way to the one that is used to evaluate the acquisition of visual targets. However, it was also found that target width plays an important role that cannot be described using Fitts' law. The final experiment, showed that when considering target width in a display a psychometric function appears when selection success rates are considered vs. target width. A successful modelling attempt of this function would result

in a uniform measure for estimating the success of the selection technique which, when considered in relation to the index of performance, can provide a simple way to evaluate the effectiveness of each interaction technique.

With respect to the fourth experiment it has to be noted, that the range of indices of difficulty was limited. In this sense, it would be interesting to design further experiments to examine the compliance of interaction speed with Fitts' law at other indices of difficulty. In particular, with respect to this issue it appears as an interesting research direction, to perform further experimentation on the applicability of Fitts' law in the non feedback case and in the case of mobile users.

In addition to the Research Questions, the thesis dealt successfully with the problem of overcoming user isolation from the real auditory environment, by using bone conductance and monaural presentation. However, issues related to bone conductance presentation and the effect of background noise on intelligibility have not been examined. It is to be expected that interference from the real world soundscape will have to be taken into account when designing the interface so that a balance is achieved where the overlapping of the two audio environments does not result in masking problems. In this sense, it would be necessary to perform further experiments to obtain design specifications for bone conductance and monaural presentation.

Another limitation of the thesis is that although spatial audio displays could be applied in a significant number of application areas, the thesis does not directly address this issue. The empirical research that was carried out, mainly addresses benchmark tasks, with people performing spatial audio target acquisition standing, with the exception of Chapter 8, where mobility is also taken into account. This was done due to the fact that no body of research on the field the thesis is investigating existed at the time of thesis development and for that reason the scope of the studies had to be contained to benchmark ones that can serve as the foundation for future studies in more complex displays and application areas.

9.5 Future Research

The future work section is again organized according to the Research Questions.

RQ 1: How can we overcome the perceptual problems in spatial audio displays and support spatial audio target acquisition?

With respect to RQ1, future research can address the problem of expanding the display area. The most important problem in this direction is the one of confusions because they will result in a ballistic movement in the wrong direction and an inefficient acquisition process. To a certain extent this problem can be alleviated using non-individualized HRTF functions and providing adequate training. Alternatively, implicit cues to the overall direction could be given using other feedback channels such as vibro-tactile stimulation or providing timbre cues. The problem of localization error can be overcome

using feedback marked spatial audio elements and making sure the display elements are separated appropriately. Future research could therefore investigate the potential of the proposed solutions in expanding the display space. This way front-back and elevation cues will become available to the designer and interaction in grid like structures using two dimensional targets will be supported.

RQ 2: What are the factors that affect deictic spatial audio target acquisition?

As was mentioned in the Limitations Section it is necessary to perform future work in the direction of display presentation and sound design for spatial audio displays. This is because aspects of display presentation can affect interaction and it is therefore necessary to find a way to balance these two display aspects. Proposals with respect to display presentation have therefore to be evaluated in order to find a presentation method that does not restrict the usability of the display.

With respect to all the experiments presented in the thesis future research could investigate the effects of practice in spatial audio target acquisition. In this way, estimations on the limits of performance for acquiring spatial audio targets will be obtained.

Gesture type was found to be a factor that significantly influences interaction performance. The three gestures that were used in Chapter 5 were successful in their own respect, however they were not examined thoroughly. In this sense it would be interesting to investigate how all these types of gestures can be integrated so that the resulting system provides multimodal input. Furthermore, physical pointing could be replaced by virtual pointers controlled by body part movements. Such gestures are more intimate and therefore could be more socially acceptable.

Feedback marked areas were found to successfully enable spatial audio target acquisition. However, different options for feedback marking the audio elements were not identified. An examination is required therefore on whether spectral or other cues can successfully substitute the feedback cue that was used in the thesis.

RQ 3: How can we evaluate the usability of deictic spatial audio target acquisition?

The thesis proposed and showed that the acquisition of feedback marked spatial audio targets can be evaluated using the techniques used for visual targets. Future work on the field could examine whether the acquisition of audio targets in the presence of different cues could be described using similar formulations. It would also be interesting to examine the acquisition of spatial audio targets in a higher range of indices of difficulty. In this way, it is possible to further examine different pointing gestures both physical as well as using virtual pointers and readily judge on their efficiency and effectiveness. The data on hand pointing could be used as benchmark ones to allow future comparisons.

The thesis also identified performance degradation for mobile participants. It is therefore of interest to investigate how this issue can be resolved using either systems that can isolate the effect of movement and filter it our or provide sufficient feedback for error recovery. Feedback for error recovery had been found useful in the case of mobile users of current handheld systems in the work of Brewster [18].

As can be seen from the future directions, the thesis has paved the way for answering a considerable number of research questions in spatial audio display research. The work in the thesis can serve as the foundation onto which future investigations in gesture interaction with a spatial audio display can be performed. Upon success, an extended spatial audio display equipped with multimodal input will provide a paradigm for shaping our future personal audio spaces.

10 APPENDIX

The companion CD ROM contains the data and the documents that were used in the experiments presented in the thesis. The data are provided in SPSS format which can be used to perform the analyses that were presented in the thesis. Here the content of each file in the CD ROM is presented.

File	Description
Files In Folder: Experimental Data & Forms\Experiment 1\	
\exp1_data\exp1_easy_comf.sav	Easy of Use and Comfort Ratings
\exp1_data\exp1_deviations.sav	Absolute Deviations from Target as a Function of Sound Location and Gesture Used
\exp1_data\exp1_up_down_results.sav	Results of the Up Down method as a Function of Sound Location and Gesture Used
\exp1_docs\exp1_evaluation.doc	The form containing the ease of use and comfort scales
\exp1_docs\exp1_partic_eval.doc	The sheet containing instructions on the informal test to informally evaluate the localization abilities of each participant
\exp1_docs\exp1_feedback.doc	Instructions for the participants that received feedback
\exp1_docs\exp1_no_feedback.doc	Instructions for the participants that did not receive feedback
\exp1_docs\exp1_information_form.doc	The experiment information form
\exp1_docs\exp1_consent_form.doc	The consent form participants had to sign
Files In Folder: Experimental Data & Forms\Experiment 2\	

\exp2_data\exp2_acc_signed_analysis.sav	Signed Deviation from Target
\exp2_data\exp2_accur_abs_analysis.sav	Unsigned Deviation from Target
\exp2_data\exp2_time_analysis.sav	Time to Complete Trials
\exp2_data\exp2_TLX.xls	Modified NASA TLX form
\exp2_docs\exp2_consent.doc	Consent & Information form
\exp2_docs\exp2_no_orientation.doc	Instructions for testing the exocentric display
\exp2_docs\exp2_orientation.doc	Instruction for testing the egocentric display
\exp2_docs\workload_estimation.doc	Instructions on Workload Estimation
Files In Folder: Experimental Data & Forms\Experiment 3\	
\exp3_data\time_one_bone.sav	Time measurements for the monaural vs. bone conductance conditions
\exp3_data\time_head_one.sav	Time measurements for the monaural vs. headphone conditions
\exp3_data\time_head_bone.sav	Time measurements for the monaural vs. headphone conditions
\exp3_data\succ_one_head.sav	Selection success rates for the monaural vs. headphone conditions
\exp3_data\succ_one_bone.sav	Selection success rates for the monaural vs. bone conductance conditions
\exp3_data\succ_bone_head.sav	Selection success rates for the bone conductance vs. headphone conditions
\exp3_data\hist_one.sav	Histogram data for the monaural case
\exp3_data\hist_bone.sav	Histogram data for the bone conductance case
\exp3_data\hist_head.sav	Histogram data for the headphone case

\exp3_data\NASA.xls	NASA TLX scores
\exp3_data\all_time_data.sav	The Time Measurements for All the Conditions in the Experiment
\exp3_docs\exp3_consent_form.doc	Consent & Information Form
\exp3_docs\exp3_instructions.doc	Instructions
\exp3_docs\NASA_TLX.doc	Modified NASA TLX form
\exp3_docs\workload_estimation.doc	Instructions on Workload Estimation
Files In Folder: Experimental Data & Forms\Experiment 4\	
\exp4_data\spss_time_standing_feedback.sav	Time Measurements for Standing Participants the received Feedback
\exp4_data\spss_time_mobile_feedback.sav	Time Measurements for Mobile Participants that received Feedback
\exp4_data\spss_time_distance_no_feedback.sav	Time as a Function of Distance to Target for participants that did not receive feedback
\exp4_data\spss_time_analysis_feedback.sav	Time Measurements for all the cases where participants received feedback
\exp4_data\spss_success_ratios_standing_no_distance.sav	Selection Success Rates for standing participants grouped over target widths
\exp4_data\spss_success_ratios_feedback.sav	Selection Success Rates for Participants that Received Feedback
\exp4_data\spss_steps_overall_no_width_all_aws.sav	Steps to select grouped per target width
\exp4_data\spss_steps_no_feedback.sav	Steps to select for participants that did not receive feedback
\exp4_data\spss_steps_mobile_feedback.sav	Steps to Select for participants that received feedback
\exp4_data\spss_steps_distance_no_feedback.sav	Steps to Select as Function of Distance to Target for participants

	that did not receive feedback
\exp4_data\spss_overall_workoload.sav	Workload Comparison
\exp4_data\spss_distance_success.sav	Selection Success Rates as a Function of Distance
\exp4_data\spss_dev_distance_no_feedback.sav	Deviation from Target as a Function of Distance
\exp4_data\overall_analysis_time.sav	Overall Time Analysis (data grouped over A/W & Target Width)
\exp4_data\overall_analysis_accuracy.sav	Overall Accuracy Analysis (data grouped over A/W & Target Width)
\exp4_data\NASA.xls	NASA TLX scores
\exp4_docs\Participant Consent Form.doc	Consent & Information Form
\exp4_docs\instructions.doc	Instructions for Participants that did not receive feedback
\exp4_docs\instructions_feedback.doc	Instructions for Participants that received feedback
\exp4_docs\NASA.doc	Modified NASA TLX form
\exp4_docs\workload_estimation.doc	Introduction to Workload Measurements

11 References

1 Accot, J. and Zhai, S. *Beyond Fitts' law: models for trajectory based HCI tasks*, in *ACM CHI*, 1997. Atlanda, Georgia, United States. p. 295-302

2 Accot, J. and Zhai, S. *Refining Fitts' law models for bivariate pointing*, in *Refining Fitts' law models for bivariate pointing*, 2003. Ft. Lauredale, Florida, USA. p. 193-200

3 Akamatsu, M., MacKenzie, S. I., and Hasbrouc, T., *A Comparison of Tactile, Auditory and Visual Feedback in a Pointing Task using a Mouse-Type device,* Ergonomics, 1995. **38**: p. 816-827.

4 Arons, B., *A Review of the Cocktail Party Effect,* Journal of the American Voice I/O Society, 1992. **12**: p. 35-50.

5 Ashmead, D., Davis, D., and Northington, A., *The contribution of listener's approaching motion to auditory distance perception,* Journal of Experimental Psychology, 1995. **21**: p. 239-256.

6 Ashmead, D., LeRoy, D., and Odom, R., *Perception of relative distances of nearby sound sources,* Perception and Psychophysics, 1990. **47**: p. 326-331.

7 Baldis, J. J. *Effects of spatial audio on memory, comprehension and preference during desktop conferences*, in *SIGCHI*, 2001. Seattle, Washington, USA. p. 166-173

8 Begault, D., *Perceptual Effects of Synthetic Reverberation on Three Dimensional Audio Systems,* Journal of Audio Engineering Society, 1992. **40**(11): p. 895-903.

9 Begault, D., *Head-up auditory displays for traffic collision avoidance system advisories: a preliminary investigation,* Human Factors, 1993. **35**(4): p. 707-717.

10 Begault, D., Wenzel, E., and Anderson, M., *Direct Comparison of the Impact of Head Tracking, Reverberation and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source,* Journal of the Audio Engineering Society, 2001. **49**(10): p. 904-916.

11 Beggs, W. D. A. and Howarth, C. I., *Movement control in a repetitive task,* Nature, 1970. **225**: p. 752-753.

12 Berkhout, A. J., de Vries, D., and Vogel, P., *Acoustic Control by Wavefield Synthesis,* Journal of the Acoustical Society of America, 1993. **93**(5): p. 2754-2778.

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

14 Blauert, J., *Spatial Hearing: The psychophysics of human sound localization*. 1999: The MIT Press.

15 Bolia, S., D'Angelo, R., and McKinley, L., *Aurally aided visual search in three-dimensional space,* Human Factors, 1999. **41**(4): p. 664-669.

16 Bregman, A., *Auditory Scene Analysis: The perceptual organization of sound*. 2000.

17 Brewster, S., *The design of sonically-enhanced widgets,* Interacting with Computers, 1998. **11(2)**: p. 211-235.

18 Brewster , S., *Overcoming the Lack of Screen Space on Mobile Computers,* Personal and Ubiquitous Computing, 2002. **6**(3): p. 188-205.

19 Brewster , S., Lumsden , J., Bell, M., Hall, M., and Tasker, S. *Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices*, in *ACM CHI*, 2003. Fort Lauderdale, FL: ACM Press, Addison-Wesley. p. 463-480

20 Brewster, S., Wright, P., and Edwards, A. *A Detailed Investigation into the Effectiveness of Earcons*, in *International Conference on Auditory Display*, 1992. Santa Fe: Addison-Wesley. p. 471-498

21 Bronkhorst, A., *Localization of real and virtual sound sources,* The Journal of the Acoustical Society of America, 1995. **98**(5): p. 2542-2553.

22 Bronkhorst, A., Veltman, J. A., and van Breda, L., *Application of a three-dimensional auditory display in a flight task.,* Human Factors, 1996. **38**(1): p. 23-33.

23 Brungart, D., Simpson, B., McKinley, R., Kordik, A., Dallman, R., and Ovenshire, D. *The Interaction between Head Tracker Latency, Source Duration and Response Time in the Localization of Virtual Sound Sources*, in *ICAD*, 2004. p.

24 Card, S. K., English, W. K., and Burr, B. J., *Evaluation of mouse, rate controlled isometric joystick, step keys and text keys for text selection on a CRT,* Ergonomics, 1978. **21**: p. 601-613.
25 Carlton, L. G., *Visual Information: The control of aiming movements,* Quarterly Journal of Experimental Psychology, 1981. **33A**: p. 87-93.
26 Chandler, D., *Semiotics for Beginners*: Routledge.
27 Cohen , M., *Throwing, pitching and catching sound: audio windowing models and modes,* Int. J. Man - Machine Studies (1993), 1993. **39**: p. 269 - 304.
28 Cohen , M. and Ludwig , L., *Multidimensional Audio Window Management,* International Journal of Man - Machine Studies, 1991. **34**: p. 319-336.
29 Cooper, M. and Petrie, H. *Three Dimensional Auditory Display: Issues in Applications for Visually Impaired Students*, in *ICAD*, 2004. Sydney, Australia. p.
30 Crispien, K., Fellbaum, C., Savidis, A., and Stephanidis, C. *A 3D-Auditory Environment for Hierarchical Navigation in Non Visual Interaction*, in *ICAD*, 1996. Palo Alto Research Center, Palo Alto. p.
31 Crossman, E. and Goodeve, P., *Feedback Control of Hand Movement and Fitts' law,* Quarterly Journal of Experimental Psychology, 1983. **35A**: p. 251-278.
32 Drury, C. G., *Application of Fitts' law to food pedal design,* Human Factors, 1975. **17**: p. 368-373.
33 Edwards, A., *Soundtrack: An Auditory Interface for Blind Users,* Human-Computer Interaction, 1989. **4**: p. 45-66.
34 Elliot, D., Hansen, S., and Mendoza, J., *Learning to Optimize Speed, Accuracy and Energy Expenditure: A Framework for Understanding Speed-Accuracy Relations in Goal Directed Aiming,* Journal of Motor Behavior, 2004. **32**(3): p. 339-351.
35 Falmagne, J.-C., *Elements of Psychophysical Theory*. Oxford Psychology Series. 1985: Oxford University Press.
36 Fitts, M. P. and Posner, I. M., *Human Performance*. 1967: Wadsworth Publishing Company.
37 Fitts, P. M., *The informational capacity of the human motor system in controlling the amplitude of movement.,* Journal of Experimental Psychology, 1954. **47**: p. 381-391.
38 Freyman, L. R., Balakrishnan, U., and Helfer, S. K., *Spatial Release from Informational Masking in Speech Recognition,* The Journal of the Acoustical Society of America, 2001. **109**(5): p. 2112-2122.
39 Friedlander, N., Schlueter, K., and Mantei, M. *Bullseye! When Fitts' Law doesn't fit*, in *ACM CHI*, 1998. Los Angeles, CA: ACM Press Addison-Wesley. p. 257-264
40 Gan, K. and Hoffman, E., *Geometrical conditions for ballistic and visually controlled movements,* Ergonomics, 1988. **31**: p. 829-829.
41 Gardner, W., *3-D Audio Using Loudspeakers*. 1997, MIT.
42 Gardner, W. and Martin, D., *HRTF measurements of a KEMAR,* Journal of the Acoustical Society of America, 1995. **97**(6): p. 3907.
43 Gaver, W. W., *The SonicFinder: An Interface That Uses Auditory Icons,* Human-Computer Interaction, 1989. **4**: p. 67-94.
44 Gilkey, R. and Anderson, T., *Binaural and Spatial Hearing in Real and Virtual Environments*. 1997: Lawrence Erlbaum Associates.
45 Gillan, D. J., Holden, K., Adam, S. R., M., and Magee, L. *How does Fitts' law fit pointing and dragging?*, in *ACM CHI*, 1990. Seattle, Washington, US. p. 227-234
46 Goose, S. and Djennane, S., *WIRE: Driving Around the Information Super-Highway,* Personal and Ubiquitous Computing, 2002. **6**(3): p. 164-175.
47 Goose, S., Kodlahalli, S., Pechter, W., and Hjelsvold, R. *Streaming speech: a framework for generating streaming 3D text-to-speech and audio presentations to wireless PDAs as specified using extensiond to SMIL*, in *International WWW Conference*, 2002. Honolulu, Hawaii, USA. p. 37.44
48 Goose, S. and Moller, C. *A 3D Audio Only Interactive Web Browser: Using Spatialization to Convey Hypermedia Document Structure*, in *7th ACM international conference on Multimedia*, 1999. Orlando, Florida, United States: ACM Press. p. 363 - 371

49 Grantham, D., *Detection and discrimination of simulated motion of auditory targets in the horizontal plane,* The Journal of the Acoustical Society of America, 1986. **79**: p. 1939-1949.

50 Grohn, M. *Localization of a Moving Virtual Sound Source in a Virtual Room, the Effect of a distracting Auditory Stimulus*, in *International Conference on Auditory Display*, 2002. Japan. p.

51 Grossman, T. and Balakrishnan, R. *Pointing at Trivariate Targets in 3D Environments*, in *CHI 2004*, 2004. Austria, Vienna. p. 447-454

52 Guiard, Y. *Disentagling relative from absolute amplitude in Fitts' law experiments*, in *ACM CHI*, 2001. Seattle, Washington. p. 315-316

53 Guiard, Y. and Beaudouin-Lafon, M., *Target acquisition in multiscale electronic worlds,* International Journal of Human-Computer Studies, 2004. **61**: p. 875-905.

54 Hart, S. G. and Wickens, C., *Workload Assessment and Prediction*, in *An approach to systems integration*. 1990, Van Nostrant Reinhold. p. 257-296.

55 Hawker, S., Soanes, C., and Spooner, A., *The Compact Oxford Dictionary, Thesaurus and Wordpower Guide*. 2002: Oxford University Press.

56 Hewlett , P., HP Official Site, http://www.hp.com

57 Holland, S., Morse, R. D., and Gedenryd, H., *AudioGPS: Spatial Audio Navigation with a Minimal Attention Interface,* Personal and Ubiquitous Computing, 2002. **6**(4): p. 253-259.

58 Intersense, Intersense Intertrax tracker Website, http://www.intersense.com/products/pro/index.htm

59 Jagacinski, R. J. and Monk, D. L., *Fitts' law in two dimensions with hand and head movements,* Journal of Motor Behavior, 1985. **17**: p. 77-95.

60 Kaernbach, C., *Adaptive Threshold Estimation with Unforced-Choice Tasks,* Journal of Perception & Psychophysics, 2001. **63**(8): p. 1377-1388.

61 Kendall, G., *A 3D sound primer by Gary Kendall. Directional Hearing and Stereo Reproduction,* Computer Music Journal, 1995. **19**(4): p. 23-46.

62 Kobayashi, M. and Schmandt, C. *Dynamic Soundscape: mapping time to space for audio browsing*, in *SIGCHI*, 1997. Atlanta, Georgia, United States: ACM Press. p. 194 - 201

63 Kramer, G., *Auditory Display: Sonification, Audification and Auditory Interfaces*. 1992, Santa Fe: Addison-Wesley Publishing Company, Inc.

64 Kvaolseth, O., *An alternative to Fitts' law,* Bulletin of the psychonomic Society, 1980. **16**: p. 371-373.

65 Langendijk, A. and Bronkhorst, W., *Fidelity of three-dimensional-sound reproduction using a virtual auditory display,* The Journal of the Acoustical Society of America, 2000. **107**(1): p. 528-537.

66 Leek, M. R., *Adaptive procedures in psychophysical research,* Journal of Perception & Psychophysics, 2001. **63**(8): p. 1279-1292.

67 Levit , H., *Transformed Up-Down methods in psychoacoustics,* The Journal of the Acoustical Society of America, 1970. **49**: p. 467-477.

68 Loomis, J., Hebert, C., and Cocinelli, J., *Active Localization of Virtual Sounds,* Journal of the Acoustical Society of America, 1990. **88(4)**: p. 1757-1764.

69 Ludwig, L., Pincever, N., and Cohen, M. *Extending the notion of a window system to audio*, in *IEEE Computer*, 1990. p. 66-72

70 Lumbreras, M. and Sanchez, J. *Interactive 3D Sound Hyperstories for Blind Children*, in *SIG CHI*, 1999. Pittsburg, PA, USA. p. 318-325

71 Lutfi, A. R., *How much masking is informational masking,* The Journal of the Acoustical Society of America, 1990. **88**(6): p. 2607-2610.

72 Lutfi, R. and Wang, W., *Correlational analysis of acoustic cues for the discrimination of auditory motion,* The Journal of the Acoustical Society of America, 1999. **106**(2): p. 919-928.

73 MacKenzie, S. and Buxton, W. *Extending Fitt's Law to Two-Dimensional Tasks.*, in *Conference on Human Factors and Computing Systems.*, 1992. Monterey, California, United States. p. 219-226

74 MacKenzie, S. I., *Fitts' law as a performance measure in human-computer interaction*. 1991, University of Toronto: Toronto.

75 MacKenzie, S. I., *Fitts' law as a Research and Design Tool in Human-Computer Interaction,* Human Computer Interaction, 1992. **7**: p. 91-139.
76 MacKenzie, S. I. and Buxton , W. *Extending Fitts' law to two-dimensional tasks*, in *ACM CHI*, 1992. New York: ACM. p. 219-226
77 MacKenzie, S. I., Sellen, A., and Buxton, W. *A Comparison of Input Devices in Elemental Pointing and Dragging Tasks*, in *Conference on Human Factors in Computing Systems*, 1991. New Orleans, Louisiana, United States: ACM Press. p.
78 Malham, D. and Myatt, A., *3-D Sound Spatialization Using Ambisonic Techniques,* Computer Music Journal, 1995. **19**(4): p. 58-70.
79 Marcus, A., *Corporate identity for iconic interface design: The graphic design perspective,* Computer Graphics and Applications, 1984. **4**(12): p. 24-32.
80 McGookin, D., *Understanding and Improving the Identification of Concurrently Presented Earcons*. 2004, University of Glasgow.
81 McGookin, D. and Brewster, S. *An Investigation into the Identification of Concurrently Presented Earcons*, in *ICAD 2003*, 2003. (Boston, MA). p. 42-46
82 Meyer, D., Abrams, R., Kornblum, S., Wright, C., and Smith, K., *Optimality in Human Motor Performance: Ideal Control of Rapid Aimed Movements,* Psychological Review, 1988. **95**(3): p. 340-370.
83 Middlebrooks, J., *Spectral Cues for Sound Localization*, in *Binaural and Spatial Listening in Real and Virtual Environments*, H.R. Gilkey and T. Anderson R., Editors. 1997, Lawrence Erlbaum Associates. p. 77-98.
84 Middlebrooks, J. and Green, D., *Sound Localization by Human Listeners,* Annual Psychology Review, 1991. **42**: p. 135-159.
85 Moore , B. C. J., *An Introduction to the Psychology of Hearing*. 3rd ed. 2001: Academic Press Limited, San Diego, CA. USA.
86 Moore, F. R., *A General Method for the Spatial Processing of Sound,* Computer Music Journal, 1983. **7**(3): p. 559-568.
87 Musicant, D. and Butler, A., *Influence of monaural spectral cues on binaural localization,* The Journal of the Acoustical Society of America, 1985. **77**(1): p. 202-208.
88 Naef, M., Staadt, O., and Gross, M. *Spatialized audio rendering for immersive virtual environments*, in *ACM symposium on Virtual reality software and technology*, 2002. Hong Kong, China: ACM. p. 65-72
89 Oldfield, S. and Parker, S., *Acuity of sound localization: a topography of auditory space. II.Pinna cues absent,* Perception, 1984. **13**: p. 601-617.
90 Oldfield, S. and Parker, S., *Acuity of sound localization: a toporaphy of auditory space. I. Normal Hearing Conditions,* Perception, 1984. **13**: p. 581-600.
91 Perrott, D., Constantino, B., and Ball, J., *Discrimination of moving events which accelerate or decelerate over the listening interval,* The Journal of the Acoustical Society of America, 1993. **93**(2): p. 1053-1057.
92 Pham, T., Schneider, G., and Goose, S. *A situated computing framework for mobile and ubiquitous multimedia access using small screen and composite devices*, in *ACM Multimedia*, 2000. Marina del Rey, California, US. p. 323-331
93 Pirhonen, A., Brewster , S., and Holguin, C. *Gestural and audio metaphors as a means of control for mobile devices*, in *ACM CHI*, 2002. Minneapolis, Minnesota, USA: ACM Press New York, NY, USA. p. 291-298
94 Polhemus, Polhemus Website, http://www.polhemus.com/
95 Pratt, J. and Abrams, A. R., *Practice and Component Submovements: The roles of Programming and Feedback in Rapid Aimed Limb Movements,* Journal of Motor Behavior, 1996. **28**(2): p. 149-157.
96 Proteau, L. and Geneviene, I., *On the Role of Visual Afferent Information for the Control of Aiming Movements Towards Targets of Different Sizes,* Journal of Motor Behavior, 2002. **34**(4): p. 367-384.

97 Pulkki, V. *Evaluating Spatial Sound with Binaural Auditory Model*, in *ICMC*, 2001. Havana, Cuba. p. 73-76
98 Raleigh, L. and Strutt, J., *Our perception of sound direction*, in *Philosophical Magazine*. 1907. p. 214-232.
99 Raman, T. V., *Auditory Interfaces: Towards the speaking computer*. 1997: Kluwer Academic Publishers.
100 Rasmussen, J., *Information Processing and Human Machine Interaction: an approach to Cognitive Engineering*. 1986, New York: Elsevier.
101 Reality, E., P5 glove Website, http://www.essentialreality.com/
102 Savidis, A., Stephanidis, C., Korte, A., Rispien, K., and Fellbaum, C. *A generic direct-manipulation 3D auditory environment for hierarchical navigation in non-visual interaction.*, in *ACM ASSETS '96*, 1996. Vancouver, Canada, 1996: ACM Press. p. 117-123
103 Sawhney, N. and Schmandt, C., *Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments,* ACM Transactions on Computer-Human Interaction, 2000. **7**(3): p. 353-383.
104 Schmandt, C. *Audio hallway: a virtual acoustic environment for browsing*, in *11th annual ACM symposium on User interface software and technology*, 1998. San Francisco, California, United States: ACM Press. p. 163 - 170
105 Schmandt, C. and Atty, M. *AudioStreamer: Exploiting Simultaneity for Listening*, in *Conference on Human Factors in Computing Systems , Mosaic of Creativity*, 1995. Denver, Colorado, United States: ACM Press. p. 218 - 219
106 Schmidt, F. L., *Statistical Significance testing and Cummulative knowledge in psychology. Implications for the Training of Researchers.,* Psychological Methods, 1996. **I**(2): p. 115-129.
107 Schmidt, R. A., Zelaznik, H. N., Hawkins, B., Frank, J. S., and Quinn, J. T., *Motot output variability,* Psychological Review, 1979. **86**: p. 415-451.
108 Sheridan, M. R., *A reappraisal of Fitts' law,* Journal of Motor Behavior, 1979(11): p. 179-188.
109 Shinn-Cunningham Barbara and Antje, I. *Selective and Divided Attention: Extracting Information from Simultaneous Sound Sources*, in *International Conference on Auditory Display*, 2004. Sydney, Australia. p.
110 Shinn-Cunningham, B., Lehnert, H., Kramer, G., Wenzel, E., and Durlach, N., *Auditory Displays*, in *Binaural & Spatial Hearing in Real & Virtual Environments*, M. Anderson R., Editor. 1997, Lawrence Baum & Associates: New Jersey, US. p. 611-663.
111 Shneiderman, B. *Direct Manipulation for Comprehensible, Predictable and Controllable User Interfaces*, in *2nd International Conference on Intelligent User Interfaces*, 1997. Orlando, Florida, United States: ACM Press. p. 33-39
112 Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 2nd ed. 1998, Reading: Addison-Wesley.
113 Shneiderman, B. and Maes, P., *Direct manipulation vs. interface agents,* interactions, 1997. **4**(6): p. 42-61.
114 Silfverberg, M., MacKenzie, S., and Korhonen, P. *Predicting Text Entry Speed on Mobile Phones*, in *ACM CHI 2000*, 2000. The Hague, Amsterdam. p. 9-16
115 So, R. H. Y., Chung, G. K. M., and Goonetilleke, R. S., *Target-Directed Head Movements in a Head-Coupled Virtual Environment: Predicting the Effects of Lag using Fitts' Law,* Human Factors, 1999. **41**(3): p. 474-485.
116 Strachan, S., Eslambolchilar, P., Murray-Smith, R., Hughes, S., and S., O. M. *GpsTunes: controlling navigation via audio feedback*, in *Mobile HCI 2005*, 2005. Salzburg, Austria. p. 275-278
117 Taylor, M. and Creelman, D., *PEST: Efficient Estimates of Probability Functions,* The Journal of the Acoustical Society of America, 1967. **41**: p. 782-787.
118 Taylor, M., Forbes, M., and Creelman, D., *PEST reduced bias in forced choice psychophysics,* The Journal of the Acoustical Society of America, 1983. **74**: p. 1367-1374.
119 Vicente, K. and Rasmussen, J., *Ecological Interface Design,* IEEE Transactions on Systems, Man and Cybernetics, 1992. **22**(4): p. 589-606.

120 Wade, M. G., Newell, K. M., and Wallace, S. A., *Decision time and movement time as a function response complexity in retarded persons,* American Journal of Mental Deficiency, 1978. **83**: p. 135-144.
121 Walker, A., Brewster, S., MacGookin, D., and Ng, A. *Diary in the Sky: A Spatial Audio Display for a Mobile Calendar*, in *IHM - HCI*, 2001. Lille, France: Springer. p.
122 Walker, A. and Brewster , S. A., *Spatial audio in small display screen devices,* Personal Technologies, 2000. **4**(2): p. 144-154.
123 Walker, B. and Lindsay, J. *Auditory Navigation Performance is Affected by Waypoint Capture Radius*, in *ICAD*, 2004. Sydney, Australia. p.
124 Wallace, S. A. and Newell, K. M., *Visual Control of Discrete Aiming Movements,* Quarterly Journal of Experimental Psychology, 1981. **35A**: p. 311-321.
125 Ware, C. and Mikaelian, H. H. *An evaluation of an eye tracker as a device for computer input*, in *SIGCHI*, 1987. p. 183-188
126 Welford, A., *The measurement of human performance,* Ergonomics, 1960. **3**: p. 189-230.
127 Welford, A. T., *The measurement of human performance,* Ergonomics, 1960. **3**: p. 189-230.
128 Wenzel, E., Arruda, M., Kistler, D., and Wightman, F., *Localization using nonindividualized head-related transfer function,* Journal of the Acoustical Society of America, 1993. **94**(1): p. 111-123.
129 Wenzel, E. and Foster, S. *Perceptual consequences of interpolating head-related transfer functions during spatial synthesis*, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1993. New Paltz,NY. p. 102-105
130 Wenzel M., E. *Effect of Increasing System Latency on Localization of Virtual Sounds*, in *Audio Engineering 16th International Conference on Spatial Sound Reproduction*, 1999. Rovaniemy, Finland: New York: Audio Engineering Society. p. 42-50
131 Wightman, F. and Kistler, D., *Headphone simulation of free-field listening,* Journal of the Acoustical Society of America, 1989. **85**(2).
132 Wightman, F. and Kistler, D., *Resolution of front-back ambiguity in spatial hearing by listener and source movement,* The Journal of the Acoustical Society of America, 1999. **105**(5): p. 2841-2853.
133 XSENS, XSENS, MT-9B, http://www.xsens.com
134 Yoshida, M., Cauraugh, J. H., and Chow, J., *Specificity of Practice, Visual Information and Intersegmental Dynamics in Rapid-Aiming Limb Movements,* Journal of Motor Behavior, 2004. **36**: p. 281-290.