Designing Intuitive Interfaces for Virtual Environments

Research Masters Thesis
Computer Vision, Image Processing, Graphics and Simulation

Christoph Grönegress
with: Mette Ramsgard Thomsen and Mel Slater

supervised by:
Professor Mel Slater

University College London, August 2001

This report is submitted as part requirement for the Research Masters degree in Computer Vision, Image Processing, Graphics and Simulation in the Department of Computer Science at University College London. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Abstract

This work investigates the feasibility of alternative means of physical interaction in virtual environments and is aimed at analysing whether humans can re-map established body functions to learn and understand to interact with digital information in a cross-sensory (virtual) environment. A further point is to attempt and establish a connection between learning, control of the interface and the degree of presence in the environment. Although the interface may not be used in work-oriented shared environments, we also wish to examine how this cross-sensory environment is received by multiple users and whether it is a factor for establishing a sense of communication and co-awareness. The application enables cross-sensory interaction by visualising certain aspects of the user's voice and, more concretely, map actions (i.e. voice) to a certain response (i.e. visualisation). In particular, we are concerned with visualising prosodic aspects of the human voice. A series of single- and multi-user studies shows that users can gain control of the intuitive interface and learn to adapt to new and previously unseen tasks in virtual environments. Though an application such as this can not be used in computer-supported collaborative work (CSCW) environments it shows great potential for arts or entertainment.

1 Introduction

1.1 Interaction

Interaction is an important aspect of our everyday life and designers of interfaces for new technologies such as virtual environments (VEs) have exploited mundane ways of interacting with digital information. Often, though, new technologies set new standards for enabling interaction, which in effect means that humans have to adapt to a wide range of user interfaces given the sheer amount of different technologies present in everyday life. In this context some people suggest that new technologies go through an early stage where preceding concepts which may not be suited for the domain dominate an interface [e.g. Oren 1990].

To ease this problem a lot of research has been carried out in designing user-friendly interfaces that allow non-experts to quickly learn to manipulate a given environment. With respect to human-virtual environment interfaces (HVEI), this means that new interface metaphors ought to be explored and used in such applications. For instance, Ishii and Ullmer [Ishii & Ullmer 1997] first bridged the gap between real space and computer interface by seamlessly combining both in the real world allowing users to grasp and manipulate digital information in a more natural way.

Another approach carried out by Slater and Usoh [Slater & Usoh 1994a] in order to link physical interaction (correctly: body-centred interaction) to a person's sense of presence in a virtual environment. Presence, which forms an integral part of current VR research, is often interpreted as the *sense of being there in VE*, and it has been suggested that it is largely defined in terms of action and interaction [e.g. Zahorik & Jenison 1998; Schumie & van der Mast 1999], though this could not be confirmed yet.

1.2 Adaptive Learning

It has been suggested that VEs show great potential for training purposes [Seidel & Chatelier 1997] and studies combining neuroscience, rehabilitation research and development in virtual environments [e.g. Rose et al 2001] suggest VR applications can be used to enforce learning in people with damaged brain functions and that this is a direct result of the adaptability and flexibility of the brain – for instance change in neural circuitry. A further implication is that a transfer to the real world does indeed take place.

If these assumptions are correct, we may find ways to exploit them in less specialised applications and design interfaces providing its users with a maximum of support for learning to control the interface.

1.3 Intuitive Interfaces

In summary of the previous sections we can define successful interfaces to exhibit the following attributes.

Ease-of-Use. User interfaces must support a high degree of flexibility in terms of user characteristics and facilitate ease of learning. Under given circumstances they also should be accessible for the disabled.

Suitability. The interface should be in harmony with the given task and the environment so as to not further complicate but support interaction. Since many applications model real-world situations, designers should model the interfaces from these and look for incorporating natural ways of interaction into their models.

Consistency. Interaction between user and the environment must be done according to a coherent scheme that maps a single action to a single outcome. An overlap between one action and multiple responses may confuse users and is therefore undesirable.

Coordination. Multiple actions carried out simultaneously should be correctly accounted for to support multi-tasking and also peripheral attention [Ishii & Ullmer 1997].

We propose that this can be realised in interfaces that support intuitive ways of interacting with environments.

As speech appears to be the major and mostly sufficient dimension of human interaction – a fact that can be demonstrated by the ubiquity of the telephone – why don't we design an interface that exploits at least parts of the human voice to establish communication between users and a digital environment?

With regards to our notes on adaptive learning, it seems a good idea to explore this in a fashion that incorporates cross-sensory interaction. Thus, mapping a number of parameters relating to one's voice into something visible and ideally graspable, could not only ease learning but also enhance the user's sense of presence; these two factors may further positively affect control and understanding of the environment.

1.4 Aims and Scope

This work is an investigation into the feasibility of alternative means of physical interaction in virtual environments and is aimed at analysing whether humans can re-map established body functions to learn and understand to

interact with digital information in a cross-sensory (virtual) environment. A further point is the attempt to establish a connection between learning, control of the interface and the degree of presence in the environment. Also, we wish to examine how this cross-sensory environment is received by multiple users and whether it is a factor for establishing a sense of communication and co-awareness.

In this respect, we shall concentrate on the design of a cross-sensory application in an immersive virtual environment (IVE) that allows us to study the points we just established. Using the voice as a source of action we will establish a number of connections between the voice and digital information in the IVE. Voice, though, should be understood as any sound made in the vocal tract that does not give rise to verbal communication directly. The term for aspects such as these is prosody and it encompasses factors such as pitch, volume, timbre etc. Therefore, don't confuse voice with speech (words and meaning). From this it should be clear that the chosen mode of *communication* strongly inhibits the natural way of communication using speech, so this study is by no means intended to support interaction relating to computer-supported collaborative work (CSCW), though the prospects of its use in other areas should be taken under scrutiny.

The application will enable cross-sensory interaction by visualising the user's voice and, more concretely, map actions (i.e. voice) to a certain response (i.e.

visualisation). In particular, we shall be concerned with visualising prosodic aspects of the human voice. The mapping should exhibit coherence as to not confuse or mislead users, but, on the other hand, it should provide users with a wide margin of possibilities.

1.5 Outline

The remainder of this thesis is structured as follows. The second and third chapter both deal with different aspects of relevant background. Chapter 2 covers necessary background on the defining attributes of virtual environments, work that has been done with regards to shared environments and interface design. Finally, we will consider presence and its relation to action and interaction, and also give an outlook on why we think our study can be successful in terms of learning. Chapter 3 states some background on speech processing and how sounds are produced in the vocal tract.

Chapter 4 briefly considers the design aspects of the virtual environment we need to build and also justifies these with respect to our analysis and experimentation, which is further covered in chapter 5, and details and considerations about experimental settings are explained there.

Chapter 6 outlines the actual application implemented during this project and considers the program structure, design issues, and means of testing.

Chapter 7 gives a thorough analysis of the data we gathered during the experimentation and we finally conclude this work by reviewing our achievements and shortcomings and also consider a few open questions to give an outlook on possible future work.

2 Background

In this chapter we will review the literature around presence, interaction and adaptive learning. We will also take a look at how cooperation among multiple users can be enforced in order to carry out various tasks and inspect whether there are guidelines aimed to improve interface design for interaction-based applications such as this one. To begin with, however, we will present some general concepts aimed at classifying virtual environments in terms of their spatial composure and other defining features.

2.1 Understanding Virtual Environments

A lot of effort has been spent on studying human experience and how interaction takes place in virtual and mixed reality environments. This not only involves assessing human factors dealing with topics such as performance and social impact of VEs, it also stresses the need for a consistent classification of different display methods and the various environments they can manifest in. A comprehensive account of these components can be useful for designing appropriate interfaces for human virtual environment interaction (HVEI).

One of the first people attempts to devise a taxonomy for VR displays was in [Robinett 1992], in which all technologically mediated experience is defined as a *synthetic experience*. The definition is deliberately kept very broad to allow different experiences such as virtual reality, teleoperation, telecommunication or even sensory prosthesis to be classified by one scheme. According to the definition, a synthetic experience can be classified in terms of a non-hierarchical nine-dimensional space: casualty, model source, time, space, superposition, display type, sensor type, action measurement type and actuator type. The first five dimensions deal with the nature and sophistication of the technology. The remaining four categories define the sensory and motor processes used in an environment. *Space*, for instance, determines how the perceived space is presented in relation to the user which can be registered with the real world, remote, miniaturised (i.e. scaled) or distorted.

Robinett's work draws some of its concepts from earlier work done by Naimark [Naimark 1991, 1992] and Zeltzer [Zeltzer 1992]. Naimark's *Elements of Realspace Imaging* organises visual experiences in terms of six methods for recording and reproducing experience. A higher category represents a greater degree of interaction and realism, with the simplest one being monoscopic imaging while the most complex one is real-time imaging which provides as much realism as unmediated experiences. Imaging techniques that lie between those two are stereoscopic imaging, multiscopic imaging, panoramics, surrogate travel, and real-time imaging.

Zeltzer's three-dimensional taxonomy is comprised of autonomy, interaction and presence – that's why it is also called the AIP cube. Whereas autonomy describes the complexity of an environment, interaction simply refers to the level of supported action, and presence is defined here in terms of the number of senses the system spans. The AIP space is defined from zero to one where a triplet defines a point in the cube; this process is similar to the colour selection process using colour cubes. A system providing perfect realism would be located at or near the point (1,1,1) while the opposite would be at (0,0,0).

Another account [Milgram & Kishino 1994], influenced by the above and other work [Sheridan 1992], describes mixed reality displays, which are defined as being located somewhere along the *virtuality continuum*, in terms of three independent factors. The virtuality continuum is bounded by the *real environment* at one end and the *virtual environment* at the other. Between these two extremes, different mixtures of virtual and real objects place the display environment somewhere along this line; augmented reality, for instance, is set in the real world enhanced by digital information. According to this definition, a mixed reality environment is therefore any hybrid environment, one that is neither purely virtual nor purely real. Furthermore, the authors suggest six classes of MR interfaces, ranging from monitor-based worlds close to the *real environment* extreme via video displays and transparent as well as non-transparent head-mounted displays (HMDs) to

completely graphical representations that are however inhabited by real physical objects.

The authors state four reasons why such a classification is not sufficient to capture the essence of all these displays. The main problem associated with them is that some are based in the real world and is enhanced with digital information, whereas others are anchored in a virtual environment enhanced with real objects. Secondly, some allow users to view the real world directly while others transmit them through some other display device (e.g. video). Thirdly, some systems are *egocentric* (i.e. immersive) while others are *exocentric* (e.g. monitor-based interfaces). Finally, accurate scale is crucial in one class and less important in another.

A decent classification therefore needs to consider a different variety of factors that also account for these differences. The solution is a three-dimensional taxonomy which encompasses the *extent of world knowledge* (EWK), *reproduction fidelity* and *extent of presence metaphor*. The first relates to the system's knowledge about the environment. This includes not only knowledge object classes that are represented but also information about their geometry, current location and the location and the pose of objects with respect to observers. A system having knowledge about a completely modelled world can correctly account for any changes in any of these parameters and needs no human supervision. Reproduction fidelity and extent

of presence metaphor both relate to the illusion of realism provided by a system. While the former is concerned with the quality with which the one can display the information – affected by hardware and rendering techniques, for instance –, the latter is about the complexity of the modelled world. All these taxonomies provide researchers with a basic framework within which they can operate. However, none of them directly addresses any issues that has to do with human behaviour or physiological capabilities. This topic is addressed next.

The need for human factors in VR interface design was first formulated by Shneiderman [Shneiderman 1992] and the desire to incorporate such knowledge into appropriate processes was expressed by Wann and Mon-Williams [Wann & Mon-Williams 1996]:

> The goal is to build virtual environments that minimize the learning required to operate within them, but maximize the information yield.

Stanney et al. [Stanney 1998] have given a thorough overview of these topics suggesting areas of human factor issues which are currently not well-understood and lack a theoretical foundation. The authors identify areas that affect human performance and also address health and safety issues and social implications of VEs. In order to maximise user performance, however,

we must consider the navigational complexity – i.e. ease or *naturalness* of navigation –, the degree of sense of presence provided by the environment[1] and user performance on benchmark tests defining a VE. A low navigational complexity coupled with a high sense of presence, and a high benchmark performance result in increasing human performance. In combination, they call for an appropriate use of VE technology that is well-suited for a given task, taking into account any inherent human sensory limitations and improving it by designing user-adaptable interfaces[2] and VE design metaphors that support action-driven user behaviour[3].

2.2 Shared Environments

2.2.1 Overview

In contrast to our work which is aimed at directly analysing how a shared environment is perceived, the main reason why shared virtual environments have been studied in the past is to support virtual meetings, computer-supported collaborative work (CSCW), for instance to effectively model crowd behaviour [Benford 1997], but also to improve multi-user computer games

[1] In this view about presence, it is rather the (objective) quality of immersive equipment than the subjective experience that determines presence.

[2] The differences of user behaviour and performance in correspondence to task and level of expertise have been studied in [Barfield & Weghorst 1993; and Dix et al. 1993].

[3] A variety of guidelines for designing VR metaphors have been suggested, for example, in [Ellis 1993; Shneiderman 1992].

[Zagal 2000] or simply to study shared environments for potential uses, for example learning [Benford 2000]. With the possible exception of computer games, all of these approaches attempt to model the real world and represent it as accurately as possible, so as to not distract users from their (worldly) tasks. Also, because of this, the environment needs to be connected to the real world in some way. We will now consider some studies in more detail.

2.2.2 Enabling Interaction

Fahlén et al. [Fahlén et al. 1993] introduce a model for enhanced user interaction based on real-world concepts of interaction and communication. The goal is to support dynamic virtual meetings, CSCW and ease of use of 'tools and services for communication and cooperation'. The concepts introduced are demonstrated by a simple application allowing users to interact with a small range of objects as well as enabling communication between users.

In this approach, individuals and functional objects are extended by an *aura*, a geometric volume (e.g. spherical) surrounding the entity. The aura is used to establish mutual awareness (or acknowledging presence) between two or more entities. Thus, in order to interact or communicate with an entity, a user's embodiment needs to be located 'inside' its aura. This definition relies on the natural concept of proximity enabling entities to interact and is

illustrated using a number of tools that can be manipulated by one or more users at the same time. Each of these have an aura, and users have to cross its border in order to be able to interact with it. A simple whiteboard, for instance, shows that many users can use it simultaneously if they are within reach. The aura of the board simply enables its tools (a drawing pen) and services (enable communication between all uses inside of aura) to whoever is inside it. Any operation carried out on the board can be seen by any other active participant. So-called documents exist that are simplified single-user versions of the whiteboard. Similarly, conference tables and podia provide a platform to extend one's aura and establish a connection to users and objects that are outside one's natural reach.

2.2.3 Collaborative Mixed Reality

Another example for supporting CSCW using Mixed Reality techniques was studied in [Billinghurst 1999] and achieved by enhancing the real world with virtual objects, which is demonstrated in two MR interfaces discussed below.

Whereas current CSCW interfaces often introduce discontinuities, or so-called seams [Ishii 1994], into the workspace – for instance by deteriorating means of communication or introducing new problem solving methods forcing users to change their traditional way of working – the authors argue that MR-based interfaces are seamless and enhance reality. Seams come in two types, one is a *functional* seam referring to *discontinuities between different workspaces*

(e.g. shared and interpersonal workspace), while a *cognitive* seam denotes any discontinuity *between existing and new practices* (e.g. computer-based and traditional desktop tools). In order to avoid seams, interfaces focussing on multi-user collaboration should therefore

- maintain traditional work techniques;
- allow users to manipulate real-world objects; and
- enable (normal) audio-visual communication.

The authors claim that interfaces taking into account these ideas will enable users to go “beyond being there” by enhancing existing techniques. Furthermore, since MR interfaces can combine virtual entities with real world input, they inherently support seamless collaboration and are therefore ideal for CSCW approaches. Indeed, Schmalstieg [Schmalstieg 1996] pointed out that the opportunity to introduce *virtuality* into the real world, the possibility of *augmentation* of objects, user *cooperation* and user (viewpoint) *independence* are natural features of any MR display. Hence, multiple users can manipulate real or virtual objects at the same time while maintaining a natural way of communication.

These ideas are verified in two MR interfaces, *WearCom* and *Collaborative Web Space*. The *WearCom* project enables multi-user conferencing and, by using transparent HMDs, allows participants to communicate with remote

collaborators while interacting with the real world at the same time. Remote collaborators appear as avatars distributed around the user.

Collaborative Web Space is a three-dimensional web browser that allows co-located users to open a number of web documents and view them at the same time. By highlighting documents users are made aware which page is currently viewed by themselves and their collaborators. Another advantage is users can enable or disable sharing of personal documents.

2.2.4 Multi-User Interaction in Computer Games

Multiplayer computer games (MPGs) are another area of interest, as they provide rules for action and interaction with other entities as well as setting up shared virtual environments; the combination of any properties can lead to a dynamic interaction between users. A model to support games design with focus on these issues is discussed in [Zagal et al. 2000]. Although the paper goes into detail about all stages of games development, ranging from idea generation to implementation, we shall only be concerned with some of the concepts relating to the characteristics of MPGs: Computer games require well-defined modes of interaction, which are categorised in two groups called the '*rules and goals'* and '*props and tools*', in this context, the following correlations are established:

Social Interaction	which defines degree of permitted interaction between two or more players. Interaction can be either stimulated or natural, i.e. permitting or prohibiting.
Competition/Cooperation	are only possible in multiplayer games, and competition refers to a type of game that determines winners and losers throughout a competitive process, while cooperative games encourage players to work towards a common goal. Both types can be combined into one game.
Synchronicity	defines the modes of action. Concurrent games require players to act simultaneously, synchronous games are turn-based
Coordination	describes the way the game process is controlled, so a single player may orchestrate a game or it may be coordinated distributed.

A small user study shows that these concepts are readily adopted by the players and that (natural) interaction among multiple users is a is an important characteristic of games.

2.2.5 Exploring the Benefits of Cooperation

By allowing children to *explore* all possibilities of multi-user environments, Benford et al. [Benford et al. 2000] focus on encouraging the users to collaborate by rewarding combined efforts with effects that can't be experienced by a single user. To analyse this, two storytelling programs have been developed (KidPad and the Klump) in which children share a single interface and use multiple input devices to draw or manipulate objects.

KidPad is a shared 2D drawing tool incorporating a zooming operation which can bring to life stories by switching between different levels of a drawing. In addition, there are a number standard of drawing tools that can be applied using a mouse. The additional features encouraging collaboration between children are intended to support tool mixing, such as two crayons can be combined by two children to generate a new colour.

The Klump is a collaborative modelling tool based around an amorphous textured polygon mesh and is designed to aid creativity in the process of story

development. It can be stretched, textured, coloured and rotated. Two or more children can manipulate the Klump at the same time. The collaborative properties of the tool enable children to pull out a larger group of vertices when two or more children combine their efforts (as opposed to a single vertex), and they can also generate a greater variety of textures when collaborating.

Informal observations in both interfaces suggest that young children are able to understand and use collaborative features of the two programs by simply exploring the opportunities. It has been found, though, that the immediate benefit of combined actions should be emphasised further, because they may otherwise be missed out.

2.2.6 Interaction across Multiple Spaces

Benford et al. [Benford et al. 1998] have constructed a shared environment that spans two real spaces and a virtual one. It induces an approach to create mixed reality spaces utilising transparent boundaries between virtual and real spaces. This is based on a classification of approaches to shared spaces in terms of transportation – the extent to which users leave behind their local space –, artificiality – the dominant domain, either real or synthetic – and spatiality – the physical properties and support for navigation in an environment. This allows us to analyse projects in terms of their feasibility (i.e.

benefits vs. cost). To illustrate these concepts and the complexity of implementing mixed reality applications, a poetry performance was staged in a virtual environment and transmitted to two distinct physical spaces, a theatre and a bar, simultaneously. The bar allowed visitors to enter the virtual environment itself via computer terminals, while both places transmitted the performance by projecting it from a fixed viewpoint onto dedicated screens. The performers were physically located on the theatre stage and used HMDs to enter the virtual stage. The project demonstrated that shared environments can be constructed spanning distinct spaces, physical and synthetic, which can provide new means for communication in VEs.

2.3 Interface Design

In this section, we will briefly review some work that has been done on interface design in order to promote interaction with virtual objects. Of particular interest here are approaches that allow users to explore the full potential of interaction while being engaged with the task of interacting.

2.3.1 Tangible Bits

In tangible bits, Ishii and Ullmer [Ishii & Ullmer 1997] present an alternative view of human computer interaction (HCI), enabling users to *grasp* and *manipulate* virtual objects that inhabit the physical world. Moving away from traditional graphical user interfaces (GUIs), the interface becomes the

surrounding (real) space enhanced with virtual objects representing the interface. Digital information is seamlessly coupled with everyday objects. Furthermore, ambient media increases awareness about peripheral processes. One application, called the *metaDESK* is aimed at shifting a computer interface from the screen into the physical environment. By using an arm-mounted display as a *window* enabling interaction with objects, users can manipulate so-called phicons (physical icons) that are projected onto a graphical surface located on a desk. Phicons represent conventional computer desktop icons that allow one to interact with an operating system.

The *ambientROOM* extends this idea to the surrounding space rather than just the desktop and is a room augmented with multimedia facilities enabling the representation of both foreground and background operations using additional ambient media such as light, shadow or sound. For instance, the number of hits of a commercial website can be transmitted using sounds and each hit is represented as a raindrop, say. The sound of heavy rain would then signify a large number of hits, while the opposite would be true for no sound. Regarding this sound as background information, users can focus on other (foreground) tasks and revert to the peripheral operation in case the sound stops or behaves unusually.

Another application is the *Tangible Geospace* which allows users to manipulate graphical maps using phicons, which are represented by physical models of landmarks of the given map. This way, one or more user(s) can

navigate, rotate, translate and scale the map by performing various tasks. The concept of tangible bits therefore successfully extends traditional interfaces to multiple senses and addresses the multi-modality of human interaction with the physical environment.

2.3.2 Interaction-Driven Interfaces

In their recent work, Parés and Parés [Parés & Parés 2001] introduce the notion of content-driven versus interaction-driven interfaces and claim that the former represent the majority of current VR application (e.g. simulations), yet they exhibit a limitation associated with exploring new interface designs don't provide a pervasive structure to support non-scientific (e.g. artistic) architectures. Before we discuss these strategies further, let's first outline the major steps in designing general VR applications:

- the simulation loop repeats until the end of the application which accounts for actions relating to user-related feedback and updates objects accordingly;
- the two-way interface maps external channels to internal ones, and also selects appropriate physical and logical or software interfaces and representations (e.g. mouse-cursor);
- object modelling and behaviour are concerned with defining objects and their behaviour according to the interaction required and to operate within the limits given by computing power of the system;

- the stimuli resulting from certain actions must be coherent with the expectations associated with these, otherwise the system may become unstable. Artistic applications, however, may ignore these conditions to establish new connections between existing mental and physical models.

A content-driven application is developed in a top-down process: first, the topic or theme is selected followed by the most suitable type of application in relation to the content. Next, the target user group is selected; this can be any grouping ranging from experts to general public. Subsequently, objects and the data are considered, which is followed by algorithmic considerations such as defining input and output devices, and the necessary modelling and development tools.

An interaction-driven strategy can be useful in certain types of applications, for instance to analyse a certain type of interface. As opposed to the content-driven approach, the interaction-driven strategy can be characterised by a bottom-up process, since we are primarily interested in defining means of interaction and their mapping. Therefore, we would begin identify the relevant input/output interfaces – the set of actions and how they are mapped –, the type of user, the type of application and its theme, until we finally consider the implementation itself (as above).

El Ball del Fanalet or *Lightpools* [see also Hoberman et al. 1999] is an interaction-driven application developed by the authors. It is a multi-user environment that is encompassed by a circular arena onto which a computer-generated image is projected. Users are given small paper lanterns (*fanalets*) each with a different artificial colour and a device that allows tracking them in the VE. A coloured spotlight is projected onto the floor at the position of each lantern and the distance from the ground is used to determine size and brightness of the spotlight; a lowered lantern results in a smaller but more intense lightpool, while a raised one produces the opposite effect. The lightpools really are *windows* that highlight parts of the virtual plane projected onto the floor. In addition, small glowing objects appear randomly for short periods of time. If they can be fed with light from a matching lantern they stabilise and grows and transforms into an articulated object. It can then be trained to dance using a lantern. The *partner* remembers combinations of movements and two users can introduce their partners to each other upon which they start dancing together using the movements they have been taught previously.

Informal observations show that most users successfully understood the environment due to an interaction metaphor that emerges during the interaction.

2.4 Presence and Interaction

2.4.1 Outline

Presence, the sense of *being there* in a virtual environment, has been studied extensively over the past decade or so[4], and it has been suggested that a higher degree of presence in a virtual environment leads to a behaviour that is similar to that in real environments under comparable conditions. Since it is about *being in a world*, either virtual or real, it has been argued [Slater et al. 1994c] that one can only speak about degree of presence in one environment compared to another.

An early account of presence was presented by Sheridan [Sheridan 1992] in which he suggested that presence is influenced by three factors, namely the fidelity of the display device, the mental illusion of accepting the environment as a real place, and basic physical interaction. In this view, presence is dependent on objective and subjective measures and in assessing it both should be accounted for [Held & Durlach 1992]. Other approaches [Slater et al. 1994b] claim that objective determinants are merely contributive rather than defining characteristic of presence while more recent accounts [Zahorik & Jenison 1998; Schubert and Regenbrecht 1999; Schumie & van der Mast 1999] tend to favour interaction as the decisive characteristic of presence.

[4] Telepresence was the first term to be mentioned by Minsky [Minsky 1980] in this context.

2.4.2 Presence in Virtual Reality

Steuer [Steuer 1992] first used presence as a defining attribute for VEs. Presence is affected by two characteristics, which, according to Steuer, are both purely stimulus-driven; that is both can be made dependent on the technical characteristics of the system. Vividness refers to technological capabilities of displaying a virtual environment while interactivity is concerned with the extent of supported actions. Both categories can be further described in terms of other qualities described below.

Vividness can be described by two factors, sensory breadth and sensory depth. The former relates to the number of senses addressed by the system (i.e. visibility, audibility, touch, smell), while the latter defines the quality of their delivery (image and sound quality). Steuer argues that it is a combination of the experience of multiple senses (high breadth) which determine presence, so, for instance, traditional media such as telephone or television are equipped with a low breadth. For a mediated experience the quality at which a system delivers must also taken into account because of limitations due to factors such as bandwidth. Breadth and depth can also affect each other in some way, such that a more restricted breadth can positively affect the perceived depth, while the same can be true for the opposite situation.

Likewise, interactivity describes the extent to which users can participate in a certain environment and can be categorised in terms of three determinants, speed, range and mapping. Speed is further defined by the system's response rate, range is determined by the number of possible actions, and mapping determines what these actions should result in. Thus, if a VR system is provided with a detailed description about these factors, it is merely dependent on the user to allow her senses to accept the current situation as a form of reality.

2.4.3 Body-Centred Interaction

In a study investigating the role of the physical body and to what extent presence is affected by physical action, Slater and Usoh [Slater & Usoh 1994a] describe the idea of *body-centred interaction* (BCI) and claim that maximising the match between the physical and virtual body when interacting with virtual entities also maximises presence. The enquiry is developed from earlier work on presence as well as sociological studies addressing the role of the human body in everyday life [Synnot 1993], according to which the physical body has several major functions enabling us to perceive, interact and communicate with other things.

An immersive virtual environment (IVE) may enhance the user's sense of presence, but it doesn't necessarily affect participants in the same way since everyone draws on different experiences and hence (slightly) different models

of the world. Focusing on the relationship between the body and the virtual body (VB), experiments suggest that the stronger the match between the sensory data and the proprioceptive sense the higher the degree of presence. In other words, the (mainly) visual input ‘prompts’ a proprioceptive ‘reaction’, and if these don’t harmonize, the sense of presence becomes unbalanced.

BCI aims at exploiting these relationships by constructing abstract models for the mapping from physical body tracking to VB dynamics by generating sensory feedback about the state of the VB in relation to the environment (e.g. shadows, reflections of the VB) and providing sufficient facilities to interact with virtual objects either *mundane* or *magical*. The crucial point is that the more body is involved in the interaction the more one feels present. The study suggests that, even though simple hand-held devices can be sufficient for navigation and interaction, presence can be enhanced significantly by employing more sophisticated mechanisms requiring users to physically interact or navigate.

The Virtual Treadmill is such an example which moves users forward whenever they walk on the spot. This is achieved through pattern matching of (image) features characteristic to walking on the spot.. The technique can be adapted to permit activities such as climbing [Slater et al. 1994b]. Experiments exploiting these mechanisms strongly support the central argument of BCI.

2.4.4 Embodied Presence

In *embodied presence*, Schubert and Regenbrecht [Schubert and Regenbrecht 1999] argue that VEs are mentally represented as *meshed patterns of actions*, and that the possibilities of physical (inter)action affect the degree of presence perceived by users.

As opposed to traditional views of presence that are based on objective measures, they suggest to put physical and cognitive activity at the centre of presence research, and propose that Glenberg's framework for *embodied cognition* [Glenberg 1997] provides the theoretical basis for it. Essentially, embodied cognition says that a cognitive representation of a scene is made up of possible patterns of actions, that captures the relationship between body and objects in a scene and results in understanding (of each part) of the environment. Understanding virtual environments can therefore be understood as the processing of mediated information and its meaning is determined as the set of possible actions in it. Furthermore, we need to suppress unwanted information from the real environment, so understanding a VE is really about suppression of stimuli and the construction and promotion of meshed patterns of actions.

2.4.5 Presence and Action

An alternative view of presence is presented in [Zahorik & Jenison 1998], which is strongly influenced by Heidegger's approach to existence in *Being and Time* [Heidegger 1993] and Gibson's ecological approach [Gibson 1979]. Instead of regarding presence as a *feeling of existence* it is rather connected to one's action in an environment, real or virtual. Presence is therefore explained from an ontological rather than a rationalistic viewpoint.

Arguing from the Heideggerian point of view that there is no subject/object divide, existence, and hence presence, should be regarded as being based on continuous action and that things in the world – and consequently the world itself – are represented in terms of their potential for action. Also, individuals are unable to reflect or predict current situations analytically; a phenomenon which Heidegger calls being-in-the-world.

Similarly, Gibson's theory of *affordance* states that an individual and its surrounding environment are related in the sense that the individual constantly picks up information from the environment and the information provided by the environment relates to possible actions.

Thus, a resulting definition of presence in terms of these approaches is straightforward:

> Presence is tantamount to successfully supported action in the environment. (Zahorik & Jenison, 1998, p.87)

Presence relates human perception to action and a virtual environment designed to support this view comprises a description of responses to dedicated user actions enables presence.

2.5 Adaptive Learning: Sensory Plasticity

2.5.1 Relevance to Present Study

Since our work is concerned with crossing senses by visualising a person's voice in different and unusual ways, thereby assessing whether it is possible for a subject to understand the relation between the own voice and the visualisation and further learn to control this new type of action, we will now take a look at a theory that, in a much larger scale, is concerned with adaptive learning of cross-sensory information.

For over 300 years, philosophers and psychologists have pondered about Molyneux's problem[5], who addressed the following question to John Locke in 1688. Suppose a blind person can learn to distinguish between spheres and cubes by touch. Now, what if this person recovers his vision? Will he be able

[5] named after the Irish philosopher William Molyneux.

to identify and distinguish between the objects visually and without the assistance of the haptic sense?[6]

Plasticity [e.g. Kolb 1995], or Sensory Plasticity [Bach-y-Rita 1967], is the field of study concerned with the compensation process of the brain in the case of lack or loss of one sense. The assumption is that the brain can undergo neuroplastic changes to enhance the sensitivity of one sensory modality (vision, hearing, touch, taste or smell) that may be used to partially compensate for the lack of another while preserving some of the key functions of the original sense. In this case, neural functions are reorganised to adapt to the different situation and a *sensory substitution* [Bach-y-Rita 1972] takes place. The brain is therefore understood as a dynamic and interactive organ that constantly changes and adapts to the stimuli it is being presented with. There is a strong relation to rehabilitation research and development in order to find ways that allow people with damaged senses to partially overcome this impairment.

Regarding the following accounts as a sideline, however not unrelated to our work, the interested reader may gain some useful insights about learning to understand cross-sensory information.

---

[6] see Degenaar's *Molyneux's Problem* [Degenaar 1996] or Morgan's *Molyneux's Question* [Morgan 1977] for an in-depth discussion of this problem.

2.5.2 Touching Text

Although the problem of reading text has been solved for blind people by using text-to-speech synthesisers, an alternative way is to *read* the text using Braille, which was developed in the 19th century by Louis Braille.

Blind individuals must somehow develop an ability to extract spatial cues from their tactile (and audio) modalities. Braille reading requires extreme sensitivity in order to translate the spatial information derived from the height and position of the dots into meaningful information. Hamilton and Pascual-Leone [Hamilton & Pascual-Leone 1998] claim that this can only be achieved by adaptive changes in the brain. This view is supported by earlier work [Recanzone et al. 1992] suggesting that the somatosensory representation of a body part (i.e. reading finger) can be altered through repeated use. Studies carried out by the authors suggest that not only the relevant section of the somatosensory cortex is enlarged, but also connections are made between this area and the occipital (visual) cortex. The authors further propose that this concept of plastic capacity may help to study other areas such as skill acquisition of normal subjects.

2.5.3 Hearing Images

Meijer [Meijer 1992] proofs the feasibility of a system that converts images into sound and reverses this process, thereby allowing blind individuals to analyse visual information using their audio sense.

The system uses greyscale images with a pixel resolution of 64 by 64 and three grey-levels, which is deemed to be sufficient to identify basic features. The mapping from image characteristics to sound are as follows: vertical position determines frequency, brightness maps to oscillation amplitude, and the column represents time. Every pixel in a column is used to generate a sinusoidal oscillator in the audible frequency range. In order to verify the accuracy of the system, an inverse function is applied and the results compared to the corresponding original images. The quality of the generated images show that image-to-sound conversion is a potentially useful description of images if potential users can learn to make sense out of the seemingly random sounds.

2.5.4 Form Perception

Bach-y-Rita et al [Bach-y-Rita 1998] have presented a way that allows blind individuals to perceive forms by stimulating their tongues using an array of electrotactile devices. This is partly motivated by the aim of building cosmetically acceptable systems that meet all the required needs.

The tongue is very sensitive to any tactile stimulus and images are transmitted to the tongue using 49-point electrotactile display originally used for fingertip stimulation. Each of the 49 electrodes can apply a small electric current which is strong enough to be noticed by a touch-sensory area such as the skin or tongue. The matrix is attached to a camera whose input is converted into electrotactile waveforms that can then be transmitted by the stimulus array.

Using 12 patterns of three geometric shapes that differ in size (i.e. 3 shapes x 4 sizes) five subjects, a study showed that performance of shape recognition is higher than for conventional skin stimulation modes.

The authors suggest that systems such as these can be useful for blind and deaf people but also for augmented communication systems such as aviation, robotics or perception in dark environments.

2.5.5 Learning in Virtual Environments

An interesting example of developing systems for people with impaired brain functions comes from [Rose et al. 2001]. They point out that VEs have been used extensively for training purposes, such as the training of fire fighters [Bliss et al. 1997], and have been found useful for psychological and physiological studies. However, more recently scientists have begun to exploit this technology to study physical disabilities [Stanton et al. 1998] and rehabilitation of people with brain damage [Brooks et all 1999].

In their studies, the authors address three problems, namely to assess:

Feasibility. The feasibility of VEs for studies with people with brain damages;

Degree of learning. whether active learning and manipulation (as opposed to passive observation) using VEs is superior to other learning techniques such as conventional presentations;

Transfer. whether a transfer from the virtual to the real world takes place, and learning be effectively enforced this way.

A series of studies with a number of (disabled) user groups[7] (i.e. people with vascular brain injury, traumatic brain injury or learning disability) suggest that VEs are indeed a good platform for studying and recover from disabilities. However, regarding more severe cases, training in VEs is not the best option. Secondly, active participation can be demonstrated to mostly result in enhanced learning compared to simple observation. Finally, a study comparing human performance of the same task or dissimilar tasks in real and virtual environments, actually shows that virtual training can be much more efficient than training in the real world, but is at least as successful as real training. Possible answers to this may be that VEs are less distractive and that learning is, in effect, more difficult in the VE and hence requires more cognitive capacity.

Concluding from this, VEs may be used as means of increasing interaction, at least for people with disabilities. Furthermore, they may give rise to changes in the brain structure of subjects and therefore be used (or abused) for targeted training of specific areas of the brain, and the studies successfully demonstrated their potential for use in rehabilitation research and development.

[7] See actual paper for further reference.

3 Speech Processing Techniques

We set out to build an environment that allows its users to manipulate objects using parts of their voice. Consequently, this chapter is dedicated to give an overview of those ideas of speech processing that are relevant to this work. In this context, we shall review some elementary phonetics and an important acoustic model of speech production: the source-filter model of speech. In addition, we will also cover some signal processing methods for speech processing and low-level recognition. It should be understood, however, that there is no reason to go into further detail here so we won't discuss what's beyond the scope of this project and concentrate on the areas stated above.

3.1 Prosody

Prosody is concerned with classifying the different qualities of a spoken sound. As opposed to phonetic transcriptions, which simply classify sounds into phonemes, however, prosodics can alter the meaning of a sequence of sounds (i.e. a phrase). Some prosodic functions are stress [Fry 1955], pitch [Halliday 1963; Ainsworth & Lindsay 1984, 1986], intensity, rhythm [Class 1939; Darwin & Donovan 1980] and duration; they can indicate the speaker's gender and age, as well as attitudes and emotional state, types of utterances

(e.g. question vs. statement), and even help analysing syntax. For instance, pitch movement – plotting pitch against time – can be used to identify the type of sentence being uttered [Halliday 1963]. In order to give a better description of prosodic terms we need an acoustic model that explains how sounds are generated and affected by the shape of the vocal tract.

3.2 The Source-Filter Model

The source-filter model of speech [Fant 1960] describes how speech is produced and affected by a number of articulators or filters. Apart from its importance in speech recognition it is also widely used in early analogue [Rosen 1958; Dennis 1962] and more recent digital [Ladefoged et al. 1978; Scully & Clark 1986] speech synthesis systems. especially formant synthesisers [Klatt 1980; Klatt & Klatt 1990].

In this model, a speech sound $s(t)$ can be correctly modelled as the convolution of a (periodic) glottal excitation sequence $e(t)$ with the impulse response of the vocal tract $v(t)$, i.e. as the convolution between a source and a filter that can vary over time. This implies that the source and filter are considered to be independent functions.

$$s(t) = e(t) * v(t) \tag{3.1}$$

Sometimes, a second filter corresponding to the lip radiation imposed on the signal is added, in which case $s(t)$ is the convolution of a source $e(t)$ convolved with a filter $v(t)$ with a filter $l(t)$.

$$s(t) = e(t) * v(t) * l(t) \tag{3.2}$$

Perceptually, a voiced sound distinguishes from an unvoiced sound in that it is generated by the periodic vibrations of the vocal cords while unvoiced sounds are produced by pure air turbulence[8]. This implies that unlike unvoiced sounds, a voiced sound is perceived as having pitch. In normal speech, the pitch of a male voice varies between 50Hz and 250Hz and for women the frequency lies between 120Hz and 500Hz.

3.3 Signal Processing for Speech Analysis

3.3.1 Overview

In this section we will analyse what common methods we can employ to pre-process a speech signal and also extract some useful parameters we can use for our interface. There are generally considered to be two approaches to signal pre-processing: parametric and non-parametric techniques. A

[8] Many sounds have a dual nature in the sense that a voiced sound intrinsically relates to an unvoiced counterpart. Examples of voiced vs. unvoiced sounds are /v/ vs. /f/, /z/ vs. /s/, /b/ vs. /p/, /d/ vs. /t/ etc.

parametric technique attempts to model speech production by approximating the parameters of the filter relating an impulse response to the actual speech output. Non-parametric techniques carry out calculations on an (isolated) *window* of the input signal, and an example of this is the Fourier transform. For a good and in-depth introduction to these techniques the reader is referred to [Cooke et al. 1993; Rosen & Howell 1991].

3.3.2 Sampling

In speech processing, we are only interested in frequencies that are in the range of human hearing – roughly between 20Hz and 20kHz – and in particular only those that adequately cover human speech sounds. In English, most of the energy lies between 2kHz and 5Khz, so a sampling rate greater than 10kHz is adequate. Regarding accuracy, a width of 12 bits is sufficient since speech signals peak at 70dB.

$$n = 70/20 \log_{10} 2 = 12 bits \qquad (3.3)$$

3.3.3 Intensity

The power of a speech signal is commonly calculated using the root-mean square (RMS) which is simply the root of the mean of the squares of the data; thus, for a discrete list $\{s_N\}$ of length N :

$$\rho = \sqrt{\frac{\sum_{t=0}^{N-1} (s[t])^2}{N}}, \qquad (3.4)$$

This is also called the volume unit or simply vu. It is strongly related to the overall perceived loudness of a signal. As we shall see below, it is also an indicator of whether a speech segment is voiced or unvoiced.

3.3.4 Voicing Analysis

Voicing is important for determining whether a signal has pitch or not [Hess 1983], and the combination of two simple techniques allow us to do so. First, consider that unvoiced sounds normally have much less energy than their voiced counterpart. Figure 3.1 (a) shows a 100ms slice of the unprocessed sound /s/, which is voiceless, while (b) shows was generated from the voiced sound /z/, also of 100ms length.

The voiced signal in (b) is much more powerful and consequently has a higher vu than (a). Thus, if the signal is above some threshold it is voiced, and below it is unvoiced. Furthermore, if the vu is below a pre-defined threshold, it is discarded and treated as non-speech, or noise.

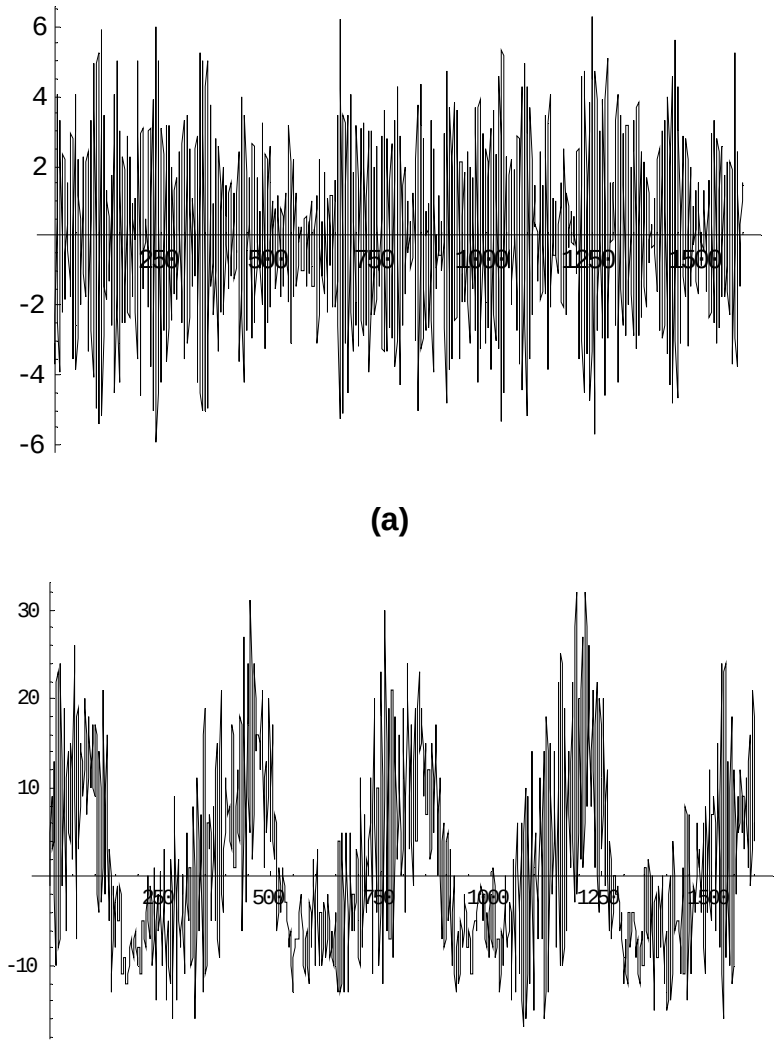


Figure 3-1: (a) 100ms of the sound /s/, vu = 2.4035, zero-crossing rate = 802. (b) 100ms of the sound /z/, vu = 9.5947, zero-crossing rate = 394.

Obviously this measure can only partly account for whether the incoming signal is voiced or unvoiced. What if the unvoiced signal, as is often the case with fricatives, contains a lot of friction, giving rise to greater energy? Figure 3.2 shows the same unvoiced signal scaled by a factor of 10. The signal has

now a vu of 24.2035 which is undoubtedly higher than that of the voiced sound above.

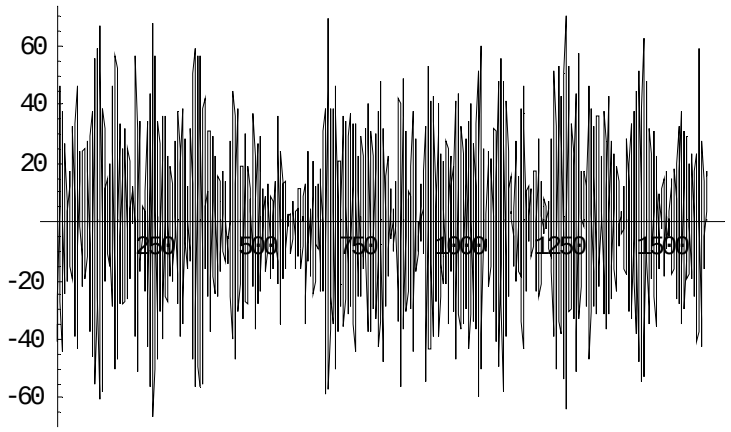


Figure 3-2: 100ms from unvoiced signal /s/, vu = 24.2035.

This shows that vu-threshold is not foolproof and does not always behave correctly in terms of determining voicing, so we either need to find a new way of measuring voicing or add another constraint. Fortunately, we can do away with the latter. One simple method, called zero-crossing, refers to the number of occurrences where the signal crosses from positive to negative or vice versa. Unvoiced speech contains a lot of high frequency friction noise, so the signal fluctuates steadily between positive and negative values and the zero-crossing rate tends to be high. This is not the case for voiced speech as the signal is excited periodically by the vocal cords and it therefore deviates from the origin for longer period. This is clearly reflected in the examples above.

While the unvoiced sound /s/ has a high number of zero-crossings (i.e. 802), there are far fewer for the voiced /z/ (i.e. 394).

To summarise, energy tends to be high in voiced sounds while zero-crossing rate is high for unvoiced sounds. We can therefore determine voicing based on a decision matrix, where the possible outcomes are illustrated in table 3-1 below, where ξ is the zero-crossing rate.

$\rho > \theta_\rho$ and $\zeta < \theta_\zeta$	voiced
$\rho < \theta_\rho$ and $\zeta > \theta_\zeta$	unvoiced
$\rho < \theta_\rho$ and $\zeta < \theta_\zeta$	noise / non-speech
$\rho > \theta_\rho$ and $\zeta > \theta_\zeta$	noise / non-speech

Table 3-1: Decision matrix for voiced/unvoiced/non-speech classification. θ_ρ and θ_ξ are the respective pre-defined thresholds for ρ and ξ.

Whereas this method is adequate to classify between voiced and unvoiced sounds, it should be noted that it is not a sufficient measure for distinguishing between speech and non-speech, but can merely make an educated guess of the occurrence of noise.

The thresholds θ_ρ and θ_ξ are best tuned for the relevant application in order to maximise performance, as these values depend heavily on the local

background noise (e.g. nearby traffic resulting in low-frequency noise around 2Hz).

3.3.5 Frequency Analysis

One method for analysing the frequency content of a periodic signal, is to compute its Fourier transform, which is an integral operator transforming a continuous function into a continuous function. Although a speech signal is not exactly periodic, voiced signals roughly exhibit the attributes of a periodic signal. This can be exemplified by simply comparing figure 3.1 (a) and (b). While (b) is voiced one can easily spot a recurring pattern, this seems impossible for the unvoiced (a). The continuous forward and inverse Fourier transforms are given by:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \cdot e^{-i\omega t} dt \tag{3.5}$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \cdot e^{-i\omega t} d\omega \tag{3.6}$$

For discrete signals, the Discrete Fourier Transform (DFT) operates on a list of length N. The forward and inverse DFT are respectively given by:

$$S[r] = \sum_{n=1}^{N-1} x[n] \cdot T^{rn} \tag{3.7}$$

$$s[n] = 1/N \sum_{n=1}^{N-1} X[r] \cdot T^{-rn}, \tag{3.8}$$

where $T = e^{-2\pi i / N}$.

A computationally less demanding algorithm was discovered by Cooley and Tukey [Cooley & Tukey 1965] and it is commonly known as the Fast Fourier Transform (FFT).

Fourier analysis can be useful in speech processing to determine the fundamental frequency $F0$, its formants and other frequency-related characteristics. For instance the spectrogram, allows one to read spectral content over time, by plotting frequency (y) against time (x). Using this technique, the formants, which are low-frequency resonances relating to $F0$, can be easily spotted in a continuous signal. This makes it easier to classify vowels, for instance.

3.3.6 Autocorrelation

Autocorrelation can be regarded as the comparison of a signal with (a delayed version of) itself and it can be used to estimate its repetition rate, the fundamental frequency, and also give an estimate of the degree of voicing of a signal. The function is expressed as the multiplication of a signal $s[n]$ with delayed version of itself $s[n+k]$.

$$\phi[k] = \sum_{n=0}^{N-1} s[n] \cdot s[n+k] \,. \qquad (3.9)$$

where k is the delay and $k \in [-N, N]$.

If the signal is voiced, $\phi[k]$ will have peaks at regular intervals, yielding an estimate of the fundamental frequency $F0$. Furthermore, local maxima will occur at the harmonics of $F0$. If, on the other hand, the spectrum is flat, the signal is noisy, which gives rise to an unvoiced or non-speech sound. From equation (3.9) it should be clear that the bigger the delay k the more computationally expensive the process becomes. This makes the use of the autocorrelation function less favourable than others, such as the voicing analysis discussed in § 3.3. Regarding pitch analysis, the method can be easily misled by background noise or confuse harmonics with fundamental frequency. It is however a very simple method that may be useful in

conjunction with other techniques. We will suggest some better approaches to pitch determination the following two sections.

3.3.7 Cepstral Processing

This technique[9] was developed by Noll [Noll 1967] and can be used to separate the excitation signal $e[t]$ from the filter $v[t]$. It allows us to accurately measure fundamental frequency and the formants of an input signal.

In the source-filter model introduced above, we assume that voiced speech $s[t]$ is the convolution of the (periodic) glottal excitation sequence $e[t]$, with the impulse response of the vocal tract $v[t]$.

$$s[n] = e[n] * v[n] \tag{3.10}$$

The convolution theorem states that time-domain convolution is equivalent to multiplication in the Fourier-domain. According to this, equation (3.10) can be represented in the Fourier domain as follows:

$$S[\omega] = E[\omega] \cdot V[\omega] \tag{3.11}$$

[9] The rather odd name cepstrum is an anagram of spectrum.

Using the logarithmic identity

$$\log(a \cdot b) = \log a + \log b \qquad (3.12)$$

we can separate the spectra of the excitation sequence $E[\omega]$ and the resonance of the vocal tract $V[\omega]$ by taking the logarithm of the above equation.

$$\log|S[\omega]| = \log|E[\omega]| + \log|V[\omega]| \qquad (3.13)$$

The cepstrum is the inverse Fourier transform of equation (3.13) and it displays the periodicities of the $E[\omega]$ and $V[\omega]$.

The *power cepstrum* reverts to the time domain and the horizontal (time) axis is called *quefruency* [Noll 1964] which is an anagram of frequency. Normally, it exhibits a number of short-duration peaks corresponding to the formants and a single long-duration peak which corresponds to the fundamental frequency $F0$. If the fundamental frequency is removed from the signal and the FFT applied once again, this results in a smoothed spectrum which allows one to detect the formants much more efficiently.

3.3.8 Pitch Determination

There exist a number of algorithms for determining pitch (or fundamental frequency) in speech signals.

These can basically be described in terms of the domain they operate in: time- or frequency-domain. The former are generally less stable and were mainly developed in the early years of speech processing. Some examples can be found in [Gruenz & Schott 1949], or [Gold & Rabiner 1969], the most successful in the past has been the method of autocorrelation described above. To verify that this method in fact works, compare the voiced and unvoiced signals in §3.3.3: the voiced signal shows signs of periodicity while the unvoiced signal appears almost random.

Frequency-domain algorithms first transform the speech input using the FFT or other frequency analysis algorithm and carry out some operations on the resulting data. If the waveform was perfectly periodic it would be simple to estimate the fundamental frequency $F0$ and its harmonics. In general though, the fundamental frequency can be estimated from the spacing of the harmonics.

One algorithm that exploits this is fact was introduced by Schroeder [Schroeder 1968]. Since the harmonic sequence in a noise-free signal is $F0,\ 2\cdot\mathrm{F0}, 3\cdot\mathrm{F0}, 4\cdot\mathrm{F0}...$, we can simply compress the spectrum successively

by the integers $2,3,4...$, and add the resulting vectors together. This yields an enhanced peak near the fundamental frequency. It should be obvious that this algorithm works without prior knowledge about the fundamental frequency, but it is very sensitive to background noise, especially high or low frequencies.

Another algorithm is based on cepstral processing presented in the last section. It is aimed at isolating a single long-duration peak which is then assigned to the fundamental frequency. Other common techniques estimate the fundamental frequency from a series of spectra see [Hermes 1993; Hess 1983] for further information.

4 Methodology

This chapter is relatively straightforward: we will justify and outline a suitable strategy and application that suits our purpose, however leaving out the implementation details until §6 where we will address this issue in some more detail. Essentially, we will determine which voice parameters we wish to use and how they will map to a certain attribute of an object and describe the environment.

4.1 Interaction with Objects

First of all, we would like to examine how people learn to interact with objects in an intuitive manner and whether being engaged in such a task can lead to a higher degree of presence. If this is the case, can it ultimately aid us in formulating a coherent theory about the characteristics of presence?

Considering the aspect of learning to interact, we cannot simply confront users with everyday objects, because one will naturally project their experiences from previous encounters to the current situation. This will not only invalidate any learning processes we are keen to observe, but it will no doubt prove to be difficult to put aside one's expectations about these objects in case they don't respond in the usual way.

Hence, in order to avoid this problem we propose to construct an abstract virtual object inhabiting an abstract space that exhibits its own behavioural pattern in correspondence to preceding user actions. This ensures that users will not get distracted by the shape and response of an object and don't necessarily associate it with some previous experience. Furthermore, we can create a new way of mapping user action (using voice) to object behaviour allowing us to maintain control over any experimental settings that might otherwise overthrow our results.

4.2 Description of the Virtual World

4.2.1 Object and Interaction

Taking into account the points we mentioned in the last section, we decided on the following option. First of all, an object appears as a wave or waveform-like entity, that in time, extends into one direction, say the negative z-axis or time axis. The object by default extends by some amount every half second (i.e. 500ms) and is represented only rudimentarily by a wireframe without any texture-mapping. It is initially flat and remains flat if there is no user action, extending only in time. A user can interact with the environment using the following (voice) parameters: voicing, volume, pitch, rhythm and silence.

Since voicing also plays a role in determining pitch and rhythm, we have three independent variables we can map to some action. For volume, pitch and rhythm, respectively, these are: width (in terms of x-axis) of an object, amplitude or height (y-axis) and stretch (z-axis). High volume results in a wider object and low volume in a narrower one. A high pitch is related to a large positive amplitude and a low pitch to a large negative one. High rate of rhythm stretches an object more while a constant tone opposes the expansion, but does not stop or reverse it. These concepts are visualised in figures (4.1), (4.2) and (4.3). In the special case of no user input (i.e. silence), the object will remain flat, but grow. Finally, we decided to completely omit voicing from the list of actions because this would result in an overlap between one action and multiple responses, which is clearly not what we want as it would then be more difficult to mentally assign one action to one response.

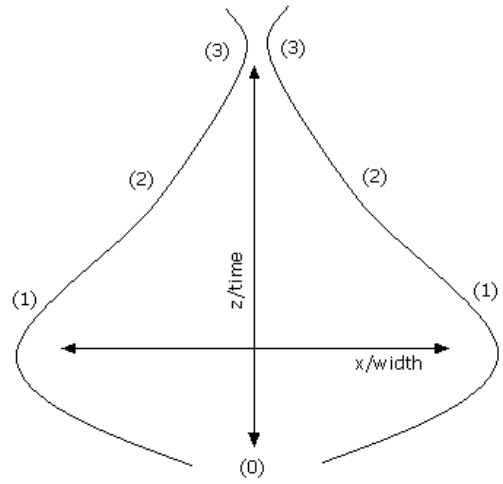


Figure 4-1: View from above/below. At (0) the object exhibits its default width. (1) corresponds to a very loud input which steadily decreases (2) and the volume is extremely low at (3).

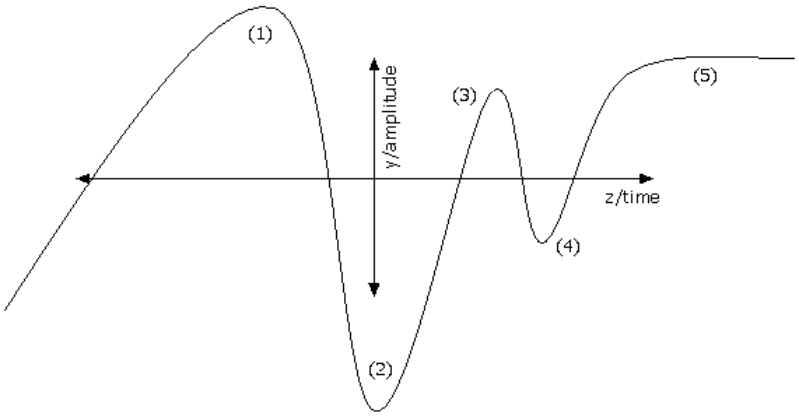


Figure 4-2: Cross-section of a waveform. (1) corresponds to a high pitch which is lowered subsequently lowered and reaches a very low tone at (2). (3) corresponds to a medium-high pitch, while (4) is the response for a medium-low pitch. (5) is a steady pitch that is higher than the mean level. A flat line parallel to the z-axis corresponds to no (pitch) input.

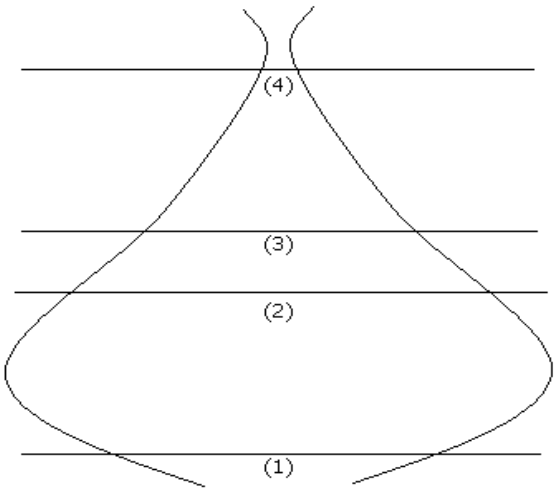


Figure 4-3: View from above/below. The axes are the same as in fig. 4-1. The horizontal lines stand for a certain time interval, e.g. 500ms. Rhythm then simply scales the length of an object between two time intervals. The rhythm having impact on the interval (1)-(2) is very high-paced, while (2)-(3) is rather flat or monotonous. (3)-(4) is again high-paced similar to (1)-(2).

Each object is equipped with a maximum length, such that it keeps on growing until it reaches its maximum upon which a new object is created at a location that is influenced by the position of the user. However, the user can further influence this and deliberately *kill* an object by remaining silent for 2 seconds, say. In this case the current object will flatten from this point onward and a new object be generated thereafter. Old objects persist for some time which is proportional to their length, or the duration the user has interacted with it. An object that has reached its maximum length due to extensive user actions, will

not only be spatially (i.e. it is longer), it will also persist for a longer while than other objects. In this sense, length of an object, as determined by user interaction, can be seen as a factor controlling its *fitness*[10].

4.2.2 Peripheral Environment and Space

To further stress the location of an object and that users can affect the location of objects, they will display a different colour for different locations. The space of objects is therefore perceived in terms of their colour.

The remaining environment is further aimed at maintaining the users attention to the objects. In fact, there exists nothing else than potentially a number of objects (though at least on will always be existent). One effect is that the space is unbounded, neither defined by some visible space nor by the natural bounds of the CAVE™. The only way the user can explore and understand the surrounding space, therefore, is by interacting with an object. A spatial understanding can therefore be constructed (only?) in conjunction with action.

[10] A term borrowed from evolutionary or genetic programming referring to the correctness (and therefore influence to later generations) of a chromosome (e.g. instructions, rules etc.) in a particular application.

4.3 Evaluation

One possible experimental scenario would be to allow subjects to freely explore the space, only giving them the clue that they can affect the environment with their voice. With regards to what we said in the beginning of this chapter, this may be a good way of dealing with things. Alternative methods would be to give subjects a specific task, for example ‘make the shape as wide/high/low as possible’. Unfortunately, we cannot guarantee that they really do (try to) understand the entire relationship this way.

An explorative task, such as the first one, might make an analysis in terms of our aims a little less complicated. So we should design our experiments and analysis around such a task. This is further dealt with in the following chapter.

5 Experimental Design

In this chapter we will progressively build up a suitable strategy to evaluate our application. We will first reiterate the hypothesis to identify the main points of interest, and then give a detailed description about design considerations such as single- and multi-user scenarios. Finally, we will select a reliable analytical tool and harmonise it with the experimental settings so as to simplify the analysis.

5.1 Preliminary Considerations

As we stated earlier, we would like to show that intuitive interaction can be learned and understood while being actively engaged with the environment. We also proposed that this may affect a person's sense of presence, ideally resulting in a higher result than for conventional *content-driven* applications. Furthermore, as a shared environment, it can facilitate the construction of a shared space between multiple users, therefore enhancing the sense of co-presence among them and a new means of communication may be established which can be regarded as an extended dimension of user communication in a virtual environment.

From this, we can see that assumptions (i) through (iii) can be studied in a single-user environment. Propositions (iv) through (vi) clearly require a multi-user scenario. Given this, it should sufficient to carry out both the single-user and the multi-user studies with a relatively small number of subjects, as our main points (i) to (iii) may possibly be correlated even though the scenarios are slightly different. We therefore propose to carry out the each of the studies with a set of around ten user groups – that is roughly ten subjects for the single-user scenario and 20 subjects, two per experiment, for the multi-user study.

In the previous chapter, we outlined the specification for a system that matches the requirements for studying these points: subjects will find themselves in an unbounded space that is populated by rectangular objects, which, in time, extend into one direction. Some of their actions will also affect an object to some extent depending on the type of action. Interaction is established through a set of intrinsic voice parameters, pitch, volume, and rhythm. These have been selected according to their following qualities: (a) they are universal parameters that can be produced without much effort, and (b) their meaning and use in reality are strongly linked to prosody in speech understanding; therefore, none of them have a ‘standalone’ implication and are only meaningful in conjunction with speech.

5.2 Single-User Study

As opposed to a multi-user set-up, the single-user study is solely focussed on interaction and presence. For this type of experiment the decision is straightforward: since there is only one user, there is no need for discussing the different voice parameters and how they can possibly be split up, so that every user can affect the objects with a different parameter of their voice. Thus, the subject will control all the parameters, that is:

Pitch	controls the 'amplitude' of a wave;
Volume	maps to width;
Rhythm	affects an object's length;
Silence	*kills* the current object and creates a new one;
User position	determines the area of birth of newly created objects.

The task can then be described as follows. A single user is sited in the CAVE™, and, in addition to the usual equipment – shutter glasses and head tracker – a wireless microphone is clipped to their shirt. The subject is then asked

> to use his or her voice to manipulate the shapes and find out what parts of their voice are important to affect the environment.

They are also told that the experiment is finished when they feel they have fully explored the space. At this stage, no further information is given to the subject about the environment or the interaction. This is because subjects are exposed to an abstract world we wish them to explore; any prior knowledge about the environment would undoubtedly alter their actions.

5.3 Multi-User Study

In addition to the single-user scenario, the multi-user set-up addresses the following questions. Can the users establish communication to each other via manipulating different aspects of an object? Is the environment experienced as a shared environment? How do the users relate to each other by manipulating the same object?

Regarding the interaction, we would ideally like to equally split the parameters though we have to bear in mind that one user is going to have a significant disadvantage for not being able to see the environment the way they really should, but rather the way it is being displayed for the person wearing the head-mounted tracker. Thus, the second user will inevitably feel less present in the environment, and it somehow seems necessary that we can somehow compensate for this by allowing them to be more engaged due to the interaction. Since we decided on allowing only two voice parameters to shape

the objects, we cannot simply grant them both to the second user or else this will impose a much greater imbalance on the possibilities for action.

Results of pilot studies suggested that subjects found it much easier to control their pitch rather than volume. In support of this view, we simply assign the volume parameter to the subject that is being tracked, while the more significant variable of pitch is allocated to the second user. Thus, the user being tracked controls the initial location of new objects and also affects them with volume; the other person is in control of pitch.

At this point, we should note that there had been the option of allowing both subjects to manipulate objects with all available parameters, and a combined effort, for instance both users uttering something at a very high pitch, could result in some amplification of a feature that could not be produced by one subject alone. The reason why we chose to split up the existing modes is because a series of single-user pilot studies suggested that users cannot comprehend their actions if there are too many possibilities; in most cases the ability to understand the mapping totally collapsed when there were more than three voice parameters linked to an object. Therefore, an extra dimension induced by an amplification would probably cause greater confusion than have positive effects on co-awareness.

The task given to the two subjects is similar to the one given to a single user. Two users are each given a wireless microphone and asked to enter the CAVE™. One of the subjects also wears the head-mounted tracker. The task given to them is

> Enter the environment and manipulate the objects with your voice and find out what actions have most impact on the environment. Also try and find out how your partner can manipulate the objects.

5.4 Analytical Methods

5.4.1 Choice of Analysis

There exist a number of methods to evaluate virtual reality studies, and they can be categorised as purely quantitative, at one extreme, or purely qualitative at the other end. One (popular) way of assessing VR studies is by using questionnaires in which subjects are asked to assign a degree of what is essentially truth or falsity to a certain question. Questionnaires are quantitative schemes and usually analysed using statistical methods, for example linear regression. This also means that questionnaires are limited to extract quantifiable information; in order to capture important information that is not quantifiable, a complementary interview can be carried out in conjunction with questionnaires.

Qualitative methods, on the other hand, aim to establish a proof using purely qualitative data derived from interviews. Since interviews make it difficult to make assumptions that are valid throughout the data set more rigorous approaches to qualitative analysis have been developed. Grounded theory is an example of this and Glaser and Strauss [Glaser & Strauss 1967; Strauss & Corbin 1990] have largely been accredited for its systematic formulation.

In general, a good strategy is to focus on one method and back the findings by use of the other.

5.4.2 Questionnaire and Interview Questions

For this study, we have agreed to put emphasis on a questionnaire-based evaluation whose results are supported by a series of substantial interviews carried out after the subjects have filled in the questionnaire. They are carried out to support the overall impressions resulting from the questionnaires but also to get hold of the data that is not easily quantifiable. Consequently, we have developed two questionnaires, one that is purely quantifiable, and another containing questions that are to be asked during an interview session that takes place after the questionnaire has been filled in by the subject.

The written questionnaire consists of three sections and 18 questions for the single-user study along with a fourth one containing four questions about

shared experiences in the multi-user study. The quantifiable questions have a range from 1 to 7, where 1 would normally indicate a total disagreement with the statement and 7 complete support of it. The first section is aimed at extracting simple demographic information such as gender, age and professional background. The second part concentrates on the interaction and how successful (if at all) the subject believes he or she achieved the task. By posing questions such as:

> After some time I controlled the objects as easily as I control my body.

and

> To what extent if at all did you feel you were in control of the objects and the environment?

This section is further directed towards the subject's perception of self and their own body image.

Most questions in the third section directly relate interaction to presence and whether or not a higher degree of interaction results in a higher degree of presence, for instance:

> The more I interacted with the objects the more present I felt in the environment.

However, most questions in this section are standard questions from existing presence questionnaires [e.g. Slater & Steed 2000].

In the multi-user study, a fourth section contains questions about co-presence, e.g.:

> To what extent if at all did you experience that you were in the same place with the other user?

and whether a means of communication was established while interacting with the same entity:

> To what extent of at all did you establish communication with the other user?

In addition to the questionnaire, we also developed a set of questions for use during the interviews. The main advantage of defining interview questions in advance is that every subject is asked exactly the same questions in exactly the same order. This ensures that the answers are correlated with respect to the question, which makes its evaluation much easier.

After a subject has filled in the questionnaire, he or she will be interviewed, during which a fixed set of questions are asked, most of which simply rephrase the questionnaire-based questions in order to substantiate the data. However, the questionnaire does not directly address the question about the virtual space and how it is perceived; we feel that it is important to gain insight

about this experience yet find it impossible to quantify such subjective a sensation and therefore base our analysis regarding spatiality purely on qualitative means. Some questions asked in relation to this are listed below.

> Where would you say the space was sited?
>
> When you were not interacting where did your sense of presence place it self?

By otherwise stating similar (and: quantifiable) questions in the questionnaire, we would presuppose only two extremes of possible sensations. Since we expect each subject to associate to the space in a more personal and different way, this would strongly impose on the actual views of the subjects.

For a full list of questionnaire and interview questions refer to Appendix A.

5.5 Protocol

The procedures for each experiment in the single- and multi-user study are very similar so to each other so it is sufficient if we a present combined description.

On entry, each subject receives information about the CAVE™ and is also instructed about any possible dangers such as epilepsy. They are then asked

to read and sign a disclaimer stating that they have been informed about a number of points. Next, they are equipped with the head-mounted tracker, a pair of shutter glasses, and a wireless microphone is attached to their shirt; in the case of a multi-user study only one subject will obviously receive a tracker. Each subject is then informed about the task and the experiment begins; the entire experiment is recorded on video. A maximum duration of 30 minutes is assigned to each experiment, but the subjects can choose to end it whenever they feel they have fully explored the space or if they become sick. After the experiment, each subject fills in a questionnaire and is finally interviewed using the set of predefined interview questions. This interview is recorded on audio tape.

6 Implementation

In the two previous chapters we outlined how are going about to gather and evaluate data for our study. This chapter outlines the software that has been implemented in order to realise this. The following sections discuss the necessary computer graphics and speech processing routines we utilised during this process.

6.1 Outline

The software needs to be capable of reading and processing speech input, while, at the same time, it has to keep track of the graphics displays of the CAVE™. We therefore need two separate processes running continuously until the program is terminated. In addition, we want the speech input to affect the graphics objects at certain intervals, so the two threads need to link up and exchange information at periodic intervals.

Essentially, the graphics thread is responsible for refreshing the four screens of the CAVE™, keeping track of the user's location and registering the left

and right eye displays[11]. Most of these issues are taken care of by CAVELib and the standard OpenGL routines.

The audio thread is in control of reading input from one or two audio channels (i.e. single- or multi-user data), and it also takes care of its processing. For this purpose, the input is split up into frames of 50ms[12], and subsequently analysed for intensity, voicing, non-speech (silence), pitch and rhythm, most of which have been described in §3.3.

The mapping from user input via parameter extraction to manipulation of the environment obviously needs to be done at regular intervals, though this procedure can be invoked by one of the main processes. Whenever the audio process has received and processed enough data, say 500ms of input, the parameters are averaged and used to update the shape – by changing the control points – and other aspects of a simple B-Spline surface. Although it is irrelevant what aspect of the shape are affected by the parameters, it should be noted that this should be done in a consistent way, so that two distinct parameters never have the same mapping. Otherwise users will never be able

[11] Since the CAVE™ wand does not play a role in the environment, it can be safely ignored.

[12] Note that for accurate speech recognition this window should be smaller than 25ms. In speech recognition, features such as short-term silent intervals are much more significant for a precise analysis. Since our software is not designed to recognise speech, and we only want to get an estimate of the fundamental frequency over a longer duration, the window size is adequate for our purposes. However, a window much bigger than 100ms would distort the actual data to a great extent.

to understand the meaning of their actions. The following sections deal with each of the threads individually.

6.2 Audio Processing

Having already introduced a number of speech processing algorithms in chapter 3, this section is fairly straightforward. All we need to do is to decide on a pitch determination algorithm and find a measure of computing a number that relates to the rhythm of an utterance. All other techniques and processes have been discussed previously and need no revision.

6.2.1 Pitch Determination Algorithms

As we outlined in §3.3, there exist a number of algorithms to track the pitch in a signal, two simple frequency-domain methods have been implemented and tested. The first is based on Schroeder's amplification of the fundamental frequency by adding to it a sequence of harmonics. Since we don't know the location of the harmonics, the spectrum is successively divided by the integers $2, 3, 4 \ldots$ and the result added to the original spectrum. Computationally, this process can be summarised as:

```
S = FFT(s)

for(n = 0; n < max; ++)
     temp = compress(S, n)
     S    = S + temp

     pitch = max(S)
```

Another pitch determination algorithm is based on cepstral processing as presented in § 3.3.6. To repeat the main steps, first compute the FFT[13] of the input signal. Obtain the logarithm of the spectrum and apply the inverse Fourier transform. The power cepstrum will exhibit a number of short-duration peaks corresponding to the formants and one long-duration peak from which we can estimate the fundamental frequency.

```
S        = FFT(s)
T        = log(abs(S))
cepstrum = inverseFFT(T)
pitch    = max(cepstrum, range)
```

Both algorithms introduced above have been implemented and were used throughout this project.

6.2.2 Rhythm

A measure of rhythm, or rate of speech, can be derived from the number of voiced/unvoiced occurrences per second. According to what we said about voicing in §3.3.4, all we need to do then is to implement two functions that compute the volume unit (vu) and the number of zero-crossings, and then tune the system to be able to distinguish between voiced and unvoiced inputs.

---

[13] For this project, we utilised a software library called the *FFTW*: *the Fastest Fourier Transform in the West*, which computes the forward and inverse DFT in one or more dimensions of real or complex data of an arbitrary size. For further information see [Frigo & Johnson 1997] or visit their website at http://www.fftw.org.

The Boolean function testing the signal for voicing is outlined below, and it returns *TRUE* if the signal is voiced and *FALSE* otherwise.

```
vu    = RMS(s)
zeroX = zeroCrossings(s)

if(vu > VU_THRESH && zeroX < 0X_THRESH)
     return TRUE

return FALSE
```

6.3 Graphics

In this section we briefly describe the tools needed to display the updated objects, how these objects are represented internally, and what data structures we utilised to make this process more efficient.

6.3.1 Representation

The objects to be manipulated are constructed from B-Spline surfaces [e.g. Ramshaw 1987]. Since objects need to be created and destroyed on-the-fly, we can employ a linked list or queue to make this process easier. The data item carried by each of the list entries contains all the necessary information about the spline, that is:

- the control points;
- the material properties;
- the transformation properties (i.e. scale, rotation, translation);

- a time stamp informing about time of creation and when it becomes obsolete (depends on user input).

Default values such as maximum length (i.e. dimensions of array of control points), the knots[14] and initial control point values are assigned globally. Note that the maximum length and the actual length (i.e. number of rows of control points) – but not necessarily the displayed length – of the spline are the same. In other words, a newly created object will force the display to only show a minimum set of rows. During every update, a counter controlling the displayed length of an object is incremented until either the user request the creation of a new one, or the counter reaches its maximum, i.e. the maximum length.

```
if(update needed)
     length++

if(length >= maxLength)
     createNewObject()
```

An object that is currently being manipulated is located at the tail of the queue, while the oldest object is at its head.

[14] For Uniform B-Splines, the knots are an evenly spaced sequence of numbers. If the knots were to be different for every object, this information would have to be local to the object and therefore combined with the individual data entry.

6.3.2 Object Creation and Destruction

New objects are created either when the current spline has reached its maximum length – which, as we discussed above, is predefined and therefore the same for every object – or if the user causes the current object to become obsolete by remaining silent for some time. One of these stopping conditions deactivates the current object and switches to a new one.

```
if(length >= maxLength || silence >= maxSilence)
     timeStamp(current)
     createNewObject()
```

In this case, the current spline is time-stamped to which the current length in rows of control points plus a default *lifetime* (in seconds) is added. The object remains displayed until this time and is removed thereafter. This ensures that objects that had been manipulated for a longer period persist longer than others.

Every frame, starting with the oldest object at the head of the queue, we check whether it needs to be removed. If so, remove it and check the next one and so on, otherwise return.

```
while(head needs to be removed)
     destroy(head)
```

When new objects are created, they are appended to the queue and equipped with a set of default values for its control points and material properties. Furthermore, a new object is placed near the current position of the user, which can be retrieved from the system using the CAVELib function

`CAVEGetPosition(CAVE_HEAD, pos)`. The surface is then active and manipulated as long as none of the two stopping conditions introduced above occurs.

6.3.3 Display

As mentioned above, most of the tasks related to the display is handled by CAVELib and OpenGL, and we only need to think about how to display all existing objects.

Now, for every object we would like to display it correctly, that is accounting for its transformation values, material properties, and shape. Since all this information is carried by each of the list elements, all we have to do is to traverse through the list of objects and display each with its individual characteristics.

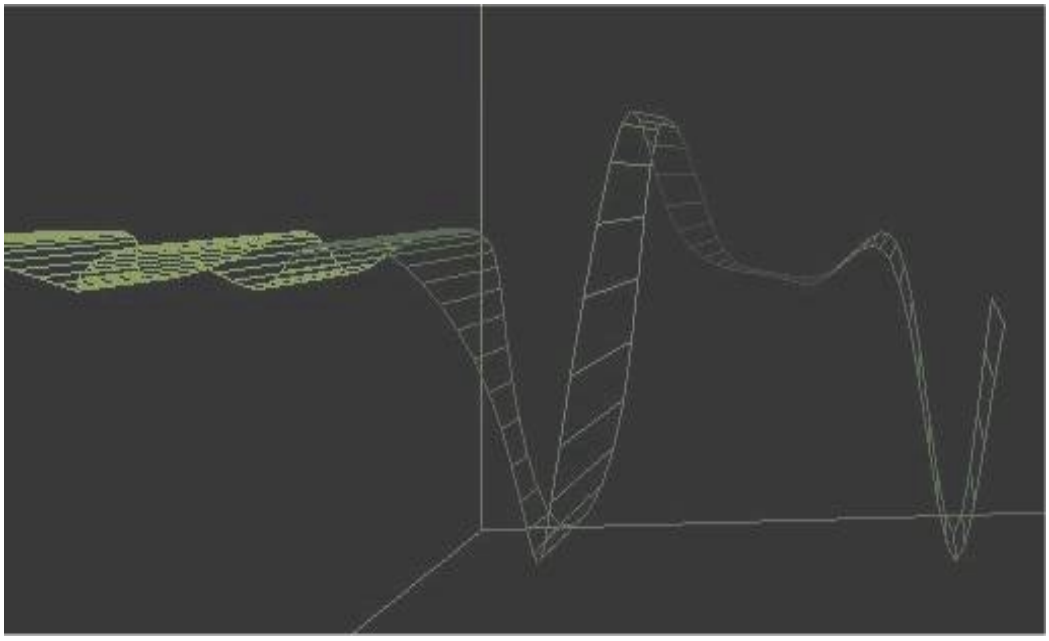
```
p = q.head

while(p)
     transform(p)
     material(p)
     drawSpline(p)
     p = p->next
```

6.3.4 Colour

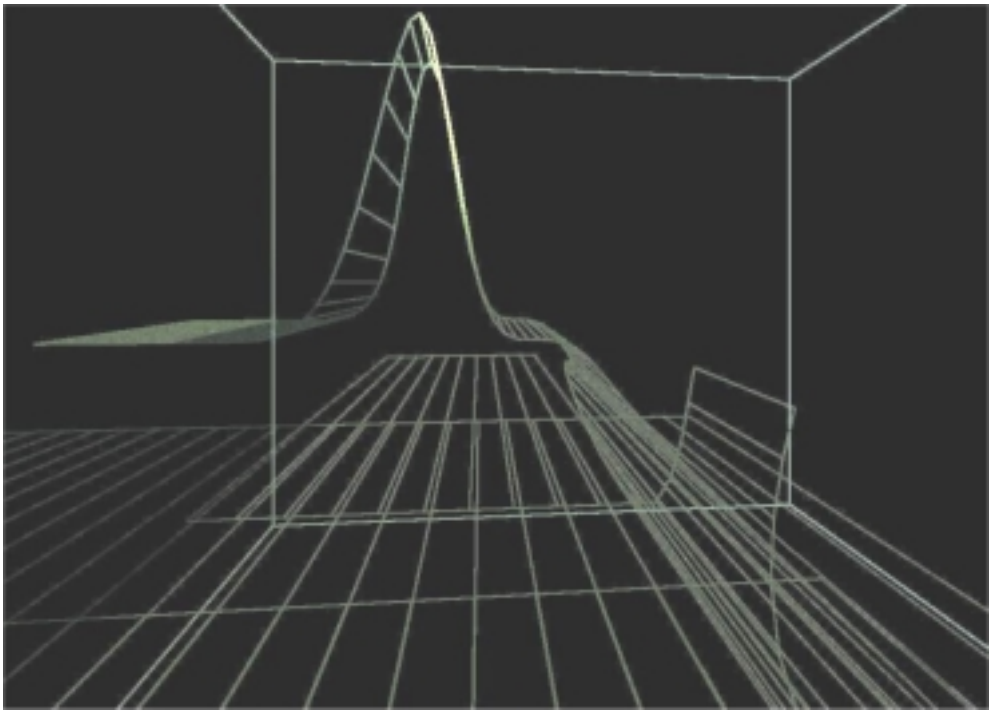
Considering colour as another parameter, we don’t directly link it to any sounds made by a subject. This is because we don’t simply want one user

action to be responsible for two or more object responses, and our repertoire of actions is limited to around five of which a few are correlated as in the case of voicing and pitch or rhythm. Alternative actions, such as using vowels or fricatives have not been implemented, so we developed another way of incorporating colour by having a fixed array of light sources whose y-axis determines the actual colour. In effect, this means that, when standing in the actual CAVE™ and facing the main wall, from left to right, the colours of the lights appear to shift from green to blue. In this sense, the position of the user determines the colour of an object whose reflectance is by default (1,1,1).

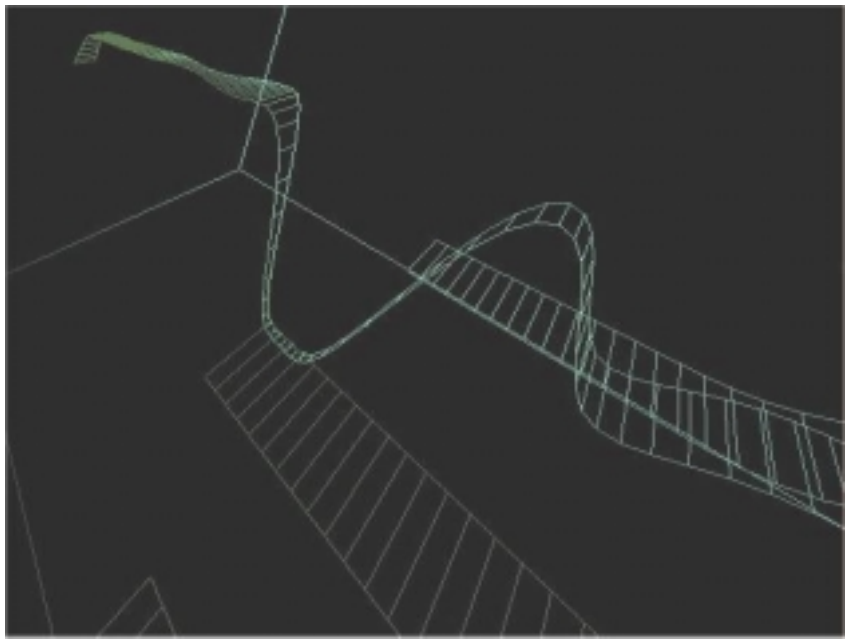
Figure 6.1 shows some snapshots taken from the CAVE™ simulator. The images mainly demonstrate the mapping of pitch, volume, silence (or stops) in a single-user environment. We found that it is extremely hard to see the impact of rhythm in two dimensions so we have omitted it here.



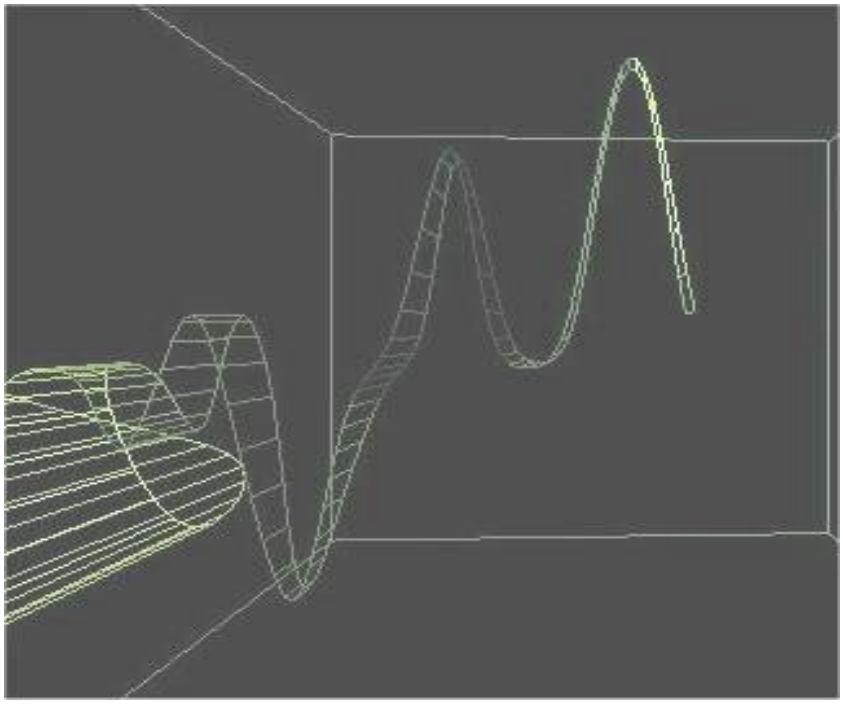
(a)



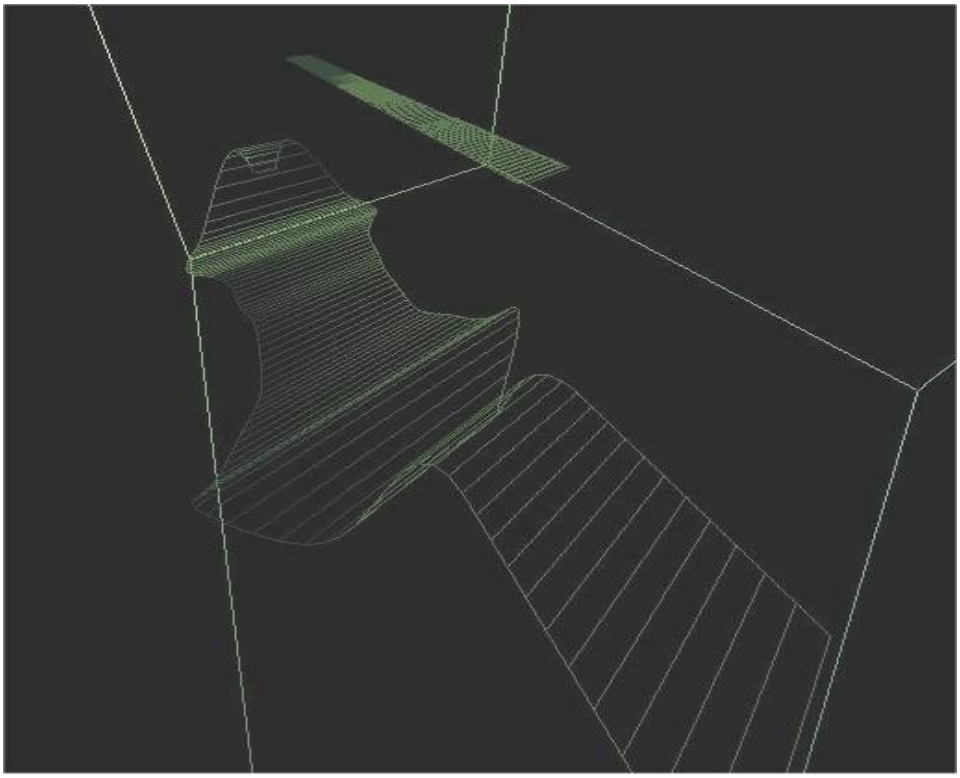
(b)



(c)



(d)



(e)

Figure 6-1: Patterns are to be read from right to left. (a) a short increasing pitch, followed by a stop (i.e. minima) followed by steady pitch. (b) back: single spike caused by a short high-pitched tone; front: flatness of object explained due to silence of user, hence one is considerably shorter than others. (c) long wave generated by a sequence of holding pitch, increasing pitch, drastically lowering pitch, while becoming more and more silent, finally increasing pitch and volume again and maintaining pitch. (d) Short very high pitch followed by loud short high pitch followed by decreasing pitch. (e) front: a moment of silence is mapped to flatness; this is followed by a very loud low pitch, which becomes louder, lower, louder, and finally pitch increases again at a lower volume.

6.4 Testing

Testing the graphics component is relatively straightforward. This is because one can clearly *see* the performance of an algorithm: in many cases, for instance, when there is something wrong with a graphics-related element, OpenGL will not display anything at all and leave the screen(s) blank instead. Hence, it is a relatively simple matter to test a new graphics component for bugs and consistency. In addition, both the graphics displayed and the complexity and degree of interaction were gradually extended throughout the project; so, while we could strongly rely on the accuracy of the existing parts, we merely had to concentrate on the new modules to find a bug.

Testing the audio processing functions, on the other hand, requires an approach a little bit more sophisticated than this. In order to verify that fundamental processes, such as, for instance, the Fast Fourier Transform, are correctly working, we compare its output with that of other data analysis tools such as Mathematica™ or IDL™. These have in-built FFT algorithms we assume to work accurately. Other algorithms were tested in a similar manner by implementing them also on a platform that provides good in-built visualisation and analysis tools.

7 Results and Evaluation

7.1 Single-User Study: Learning, Presence and Interaction

For the single-user study there were a total of eight subjects, with the majority of them male (6) and an average age of 34.5 years. All subjects exhibited a surprisingly high degree of presence given the abstract environment; the scores from the questionnaire result in a mean presence of 5.661 (median: 5.667), while the extent of interaction was also confirmed strongly with a mean of 5.125 (median: 6). The extent to which subjects achieved their tasks was rated a little lower with a mean of 4 (median: 4), and so was the sense that subjects actually were in control, mean: 3.75, median: 3.5.

In this respect, the results are a somewhat twofold, not necessarily inconsistent though: whereas subjects did report a very high sense of presence and interaction, this could not directly be correlated with the success of learning to control the environment. At the same time, however, the sense of identification with the objects was relatively high at a mean level of 5 (median: 5) from which we assume that some kind of relation between user and environment was nonetheless established.

Taking a closer look at the data, we can first direct our attention at how age plays a role in achievement of the task. These data are related in the scatter diagram below (7-1). Though there are maybe two outliers, the linear regression clearly shows that older subjects had more problems with the environment than younger ones.

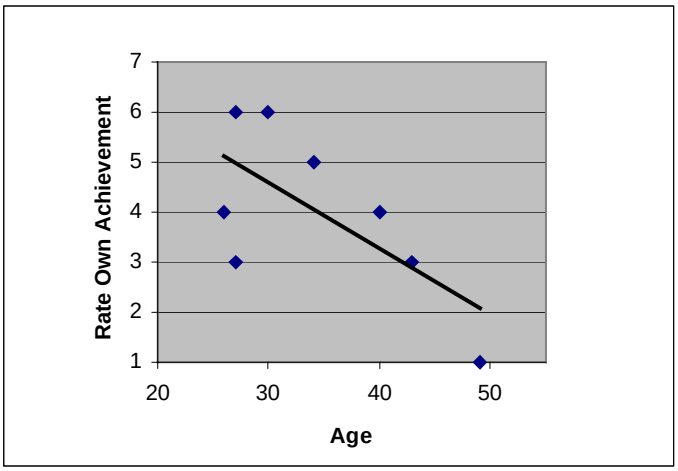


Diagram 7-1: A higher age seems to negatively affect a subject's performance and ability to understand the environment.

Regarding successful learning in general, we can make further assumptions. First of all, the data should confer that a higher achievement also means higher control. This is indeed the case as diagram (7-2) illustrates. A linear regression of the data points shows the relation between achievement and control. This is further supported by connecting achievement with the degree to which objects were linked to one's voice as is done in (7-3).

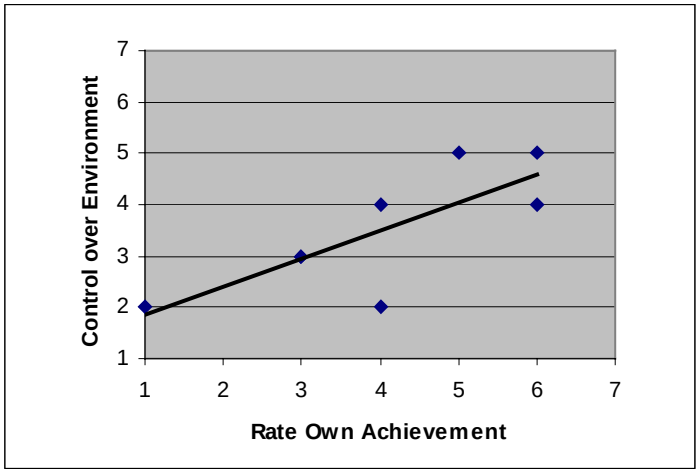


Diagram 7-2: Linear regression of the diagram verifies that the data are correlated.

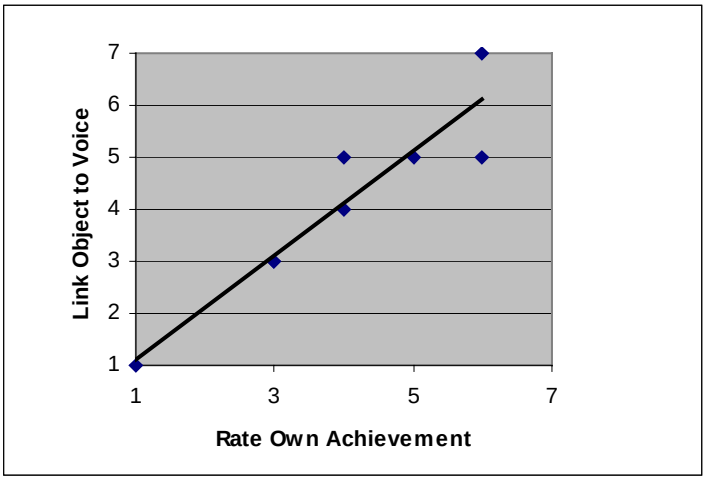


Diagram 7-3: Successful task completion is linked to a subject's voice.

Next we shall see how the length of an experiment affected the degree of interaction; diagram (7-4) refers to this relation. It demonstrates quite effectively that there seems to be a minimum amount of time a person has to

spend in the environment in order to appreciate it. In this context, one subject spoke of two different experiences, where he assigned the first half a value of 3 that was centred around 6 in the end. Two other subjects noted:

> Much stronger sense of environment towards end.

and

> When I stopped to interact I almost had the feeling of loneliness because the forms would not do anything then.

While the first statement demonstrates that subjects did undergo a learning process, the second one could be interpreted as stating that, the way the environment responded to the lack of action, it encouraged users to become more engaged and increase the interaction with it, which consequently seemed to be rewarded in some sense. Both subjects that spent around ten minutes in the environment had very little control over the environment, i.e. 2 and 3, so time may be affecting control, though this view could not be verified by the data.

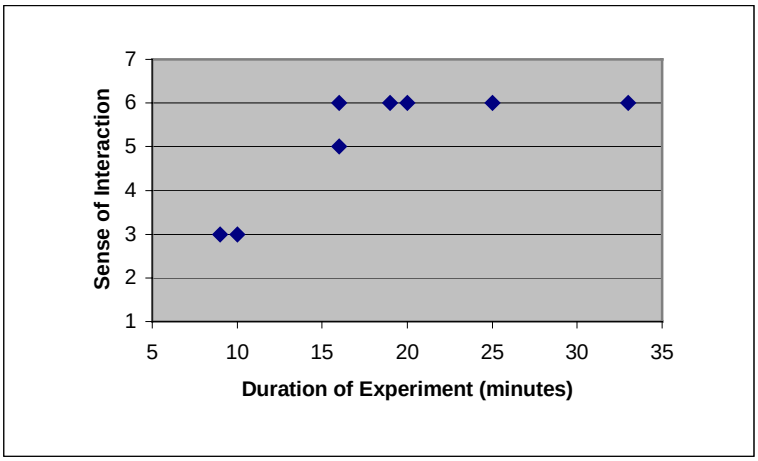


Diagram 7-4: Length of experiments against degree of interaction. There seems to be a threshold, between 10 and 15 minutes, below which interaction appears meaningless to a subject while it is better understood after a longer duration of experimenting with the environment.

From the interviews we gained the following insight that most subjects' perception of the sense of presence was twofold: whereas in the beginning, most people reported a particularly low sense of presence (numerically between 2 and 3), this sensation seemed to gradually increase over time and most subjects reported a similar experience for the sense of interaction. Interaction was inherently linked to a subject's sense of presence which can be verified by scores given to the question linking these in the written assessment (mean: 6.125, median: 6). We did not further pursue the question of why people felt less present in the beginning compared to the end, but speculate that this might be connected to control over the environment,

although the data does not show this, because all subjects, including those who failed to achieve to gain control, reported a very high sense of presence.

We can conclude from this that learning does indeed take place and the younger the subject the more likely they are to achieve this. Also, quite logically, a longer duration of learning results in a deeper understanding of the interaction and a greater learning. Presence can be linked to interaction, although this could not verified by the statistical data itself, it seems plausible by correlating the statistical data and the answers given in the interview.

What, in addition to this, caused more concern was the high degree of presence of all the participating subjects because we could not explain it at all and we can offer two explanations. A simple answer would be to claim that one or two occurrences that are incoherent are, in fact, outliers and should be regarded as such. Another possibility however is that the nature of the abstract environment forced users to either accept it as their dominant reality in order to be in a position to interact with it at all, or ignore it, in which case users could ‘by default’ not interact with their environment. This may imply that interaction is linked to presence in the sense that (inter)action is only possible or feasible in an environment that one feels present in. This thought is purely hypothetical, though, but it would be interesting to consider it in another experiment. The relation between sense of presence and interaction is shown

in figure 7-5 below. Although the data is very tightly spread, there is a definite trend from strong link and high presence to weaker link and lower presence.

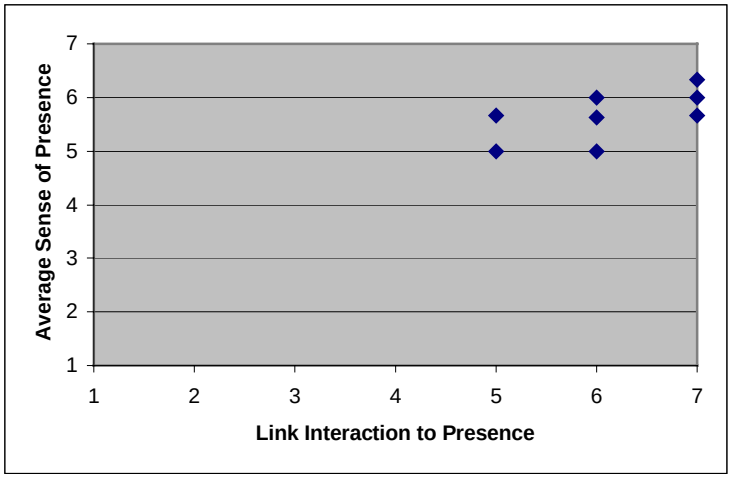


Figure 7-5: Relation between the link between interaction and presence to the average sense of presence.

Regarding the mental construction of a virtual space, the unbounded space seemed to have various effects on subjects. Although this was not part of the written questionnaire, we did confront subjects with this question during the interview. Asking them *where* they thought the space was sited, two subjects said:

> A computer world. I felt like I was inside a computer.

and

> It was neither real nor unreal, because I was there. Some kind of extra dimension.

Towards the question of size of the object most subjects answered ‘infinite’ or ‘unbounded’, though one subject said three by three meters (the dimensions of the CAVE™). As a final note, it should be interesting to mention that two users ran into one of the display walls on at least one occasion, so the belief that the space was in effect perceived as much greater than it actually is can be impressively demonstrated by this.

7.2 Multi-User Study

During the multi-user study, we encountered a number of problems, which ultimately led us to abandon and not further pursuit this part of the project after three experiments had been carried out. The problems we encountered had two sources:

- software/hardware disruptions;
- social interactions between members of the same group.

Regarding the former, whereas the quality of the microphones was sufficient to prevent too much noise from contaminating the speech input in the single-user environment, they had not sufficed to deal with one user affecting the (microphone) input of the other and vice versa. It turned out that this essentially resulted in noise – and therefore useless data – in both channels. Alternatively, we tried to alleviate this by using a number of simple filtering

algorithms, but that also failed to improve matters. At that point, we could have continued our experiments, knowing that user would never be able to control the environment at all. Thus, we could have limited our analysis to user interaction and co-presence, but there was another factor that led us to withdraw from the study.

Whereas we could (and should!) have foreseen this first problem arising – due to pilot studies for instance – the second one was effectively beyond our control and left us utterly puzzled: the fact that the subjects were introduced to each other for the first time just before the experiment, not knowing each other beforehand, resulted in various peculiar situations and also seemed to change people's focus from a task-oriented performance to a more "user-oriented" one.

During one experiment, for example, this simply resulted in two users seemingly entering into a competition as to who could speak with greater force. Whether they decided to do so because they wanted to prevent instead of assist each other from understanding the space – and therefore work in an indisputably counterproductive way – or whether both genuinely thought, this would be a decent heuristic approach is not clear. The experiment finished without any connection being made and one of the subjects, who had obviously *lost* the competition by that time, stood silently in one corner of the

CAVE™ and watched the other trying to interact. In the subsequent interview, this subject, who incidentally was wearing the head tracker, also noted:

> Having two people made it more of a competition than an interaction.

Although, with an average of 5.25, he awarded relatively high scores to questions about co-presence and communication to the user, he admitted:

> I guess I'm more inhibited than I [thought] but another person in the environment helped to remove those.

So, this expression may partially account for this subject's high scores on co-presence, as those of the second one resulted in an average of 2.

In another experiment, two users started with a promising conversation about how they would go about solving the task. Unfortunately though, they soon got distracted, admittedly due to the failure of the system, and kept on chatting about something unrelated to the task instead. This group also failed to achieve any control, though at least their scores on co-presence were somewhat related (4.25 and 5.75) and their sense of presence shows a similar relation (3.334 and 4.667). One of these subjects remarked:

> I enjoyed the experience, however I feel I did not achieve what I was suppose[d] to as I didn't feel my voice changed the shapes.

The questions we have to ask ourselves are: would this have happened if the system would have been made to work accurately? Otherwise, how can disturbances like these so drastically sabotage the outcome of an experiment? Much more importantly, though, is there anything we can do in preparation to eliminate such incidences by default? Why do multi-user studies cause so much more trouble than single-user studies when the task and the environment (at least visually) in both scenarios are practically the same?

Although a comparative study of group behaviour in real and virtual environments has already addressed related problems [Slater et al 1998], the impact of social behaviour towards the outcome of (VR) experiments certainly needs to be further investigated. We will further address this problem in the next chapter.

8 Conclusions

In this work, we presented a new and cross-sensory way of interacting with virtual objects by employing intuitive characteristics of the human voice. We have visualised parameters of the human voice in a 3D immersive virtual environment and shown that users can gain control of the intuitive interface and learn to adapt to new and previously unseen tasks in virtual environments. We failed, however, to convincingly link successful control of an environment to a higher degree of presence, which, in our case, seemed to have been supported by the exceptional visual experience more than anything else. Nonetheless, we could associate a high degree of presence with a high degree of interaction, which seems to at least partially support the recent trend towards explaining presence through interaction (see §2.4 for details).

Even though we fell short to demonstrate any assumptions regarding shared spaces (the reasons for which were discussed in the previous chapter), we gained a lot of insight about the impact of group behaviour on the actual experiments. This may have implications on the conduct of any future multi-user study, because it shows that there are matters that are currently not being controlled, measured or accounted for in the design of environments or conduct of standard experiments. Subjects are ultimately more important to

the success of an experiment than these factors, so we need to find a way to incorporate more knowledge about human behaviour in social groups into our models that might help and overcome these problems.

Even though we experienced technical faults with our equipment during the multi-user study, one or two related scenarios may be worthwhile to examine: multiple users could interact with the same information across different physical spaces, either in a Mixed Reality situation or a study that links two CAVEs™ (or other) via a network connection. The problem of two users affecting each other's input channel would be avoided, and either situation may be interesting to pursue as any question relating to shared space, co-presence and communication would be fundamentally different to the present work.

Regarding both studies, it seemed as if we were unable to draw any representative conclusions from any of them, because there weren't enough subjects that allowed us to prove or disprove an assumption. Regarding the use of analytical tools, qualitative methods might be more suitable and insightful with respect to the kind of interface than quantitative ones. Future studies in this area should consider these two factors.

Moreover, in the case of a further pursuit of the benefits of intuitive interaction one must consider that *degree of interaction* may under some circumstances

be related to one's shyness, which is certainly true for our study. So, if a person's nature has an impact on the outcome of a study, it should be accounted for in terms of an analysis. Shy people, for instance, would naturally dislike to (inter)act with things explicitly. Although we were aware of this problem, we found no rigorous method of controlling it. Future studies in intuitive interaction must therefore find a way of dealing with this problem and at least incorporate questionnaires that can explain, for instance, a low degree of interaction with shyness.

Finally, there are many open questions regarding the benefits of using voice as a means of interaction in VEs. Throughout our studies, we deliberately avoided using more than three voice parameters to map to an object. We found that even three parameters produced by the same organ at the same time were almost too much to comprehend, which is also one explanation for why the scores for control failed to impress. Even if the user concentrates on a single aspect of the interaction, it is always the full range that is interpreted, which makes it harder for him to understand the mapping. In this sense, the voice may not be a very good medium to use for interaction with digital information as – at least in our application – users normally control three actions (pitch, volume, rhythm) at a time. Also, regarding conventional multi-user environments, the voice is normally assigned to the – quite meaningful – task of verbal communication between participants, so how could parts of it be used to additionally manipulate the environment? This almost seems

impossible not to say meaningless. However, cross-sensory interfaces like the one described here bear great potential for use in arts or entertainment as almost every subject reported they had enjoyed the experiment and the environment a great deal.

Acknowledgements

This project was partially funded by the Equator EPSRC project.

Thanks also to David Swapp for giving advice on implementation-related issues.

Appendix A: Questionnaires and Interview Questions

The multi-user questionnaire presented in A.1 contains the entire single-user questionnaire which uses questions 1 to 17 and the last question in addition to four more questions on shared experience and co-presence. For further detail about design considerations refer to §5.

The interview questions printed in A.2 has a single section on co-presence while the remainder has been used throughout both studies.

A.2 Written Questionnaire

1. What is your gender?

male female

2. What is your age?

3. What is your professional background ?

4. I have experienced virtual reality

1 2 3 4 5 6 7

never a great deal

5. How much time did you spend in the environment?

6. I achieved my tasks

1 2 3 4 5 6 7

not very well very well

7. How strong was your sense of interaction with the environment?

1 2 3 4 5 6 7

not at all very strong

8. To what extent if at all did you feel you were in control of the objects and the environment?

1 2 3 4 5 6 7

not at all nearly all of the time

9. After some time I controlled the objects as easily as I control my body.

1 2 3 4 5 6 7

not at all very much so

10. Passively viewing the environment develop did not make a difference from getting actively involved.

1 2 3 4 5 6 7

not true very true

11.

(a) During your experience which of the following actions had any impact on the environment? Please mark any applicable and specify where requested.

I don't know	using the head-mounted tracker	words I uttered (please specify)	intrinsic speech parameters, e.g. intonation, pitch. (please specify)	movement/ physical location (please specify)

(b) To what extent did you utilise any of these actions in an unusual/unnatural manner in order to affect the environment?

1	2	3	4	5	6	7
not at all						a great deal

12. To what extent if at all did you relate to or identify with the objects?

1	2	3	4	5	6	7
never						a great deal

13. To what extent did you feel the objects were linked to your voice?

1	2	3	4	5	6	7
not at all						a great deal

14. To what extent if at all did you feel that there were times at which the environment became the dominant reality for you and you almost forgot about the surroundings of the lab?

1	2	3	4	5	6	7
never						almost all of the time

15. To what extent if at all did you feel present in the environment, where 7 represents your *normal experience of being in a place.*

1	2	3	4	5	6	7
not at all						very much so

16. During the time of the experience which was the strongest on the whole, your experience of being in the environment or of being in the lab?

1	2	3	4	5	6	7
in the lab						in the environment

17. The more I interacted with the objects the more present I felt in the environment.

1	2	3	4	5	6	7
not at all true						very true

18. To what extent if at all did you experience that you were in the same place?

1	2	3	4	5	6	7
not at all						a great deal

19. To what extent if at all did the other user have an impact on the forms?

1	2	3	4	5	6	7
not at all						a great deal

20. To what extent of at all did you establish communication with the other user?

1	2	3	4	5	6	7
not at all						a great deal

21. To what extent were you and the other person in harmony during the course of the performance of the task?

1	2	3	4	5	6	7
not at all						a great deal

22. Any further comments (write on back of paper if necessary):

A.2 Interview Questions

1) How would you describe your sensation of interaction with the installation?

2) During the time of the experiment how did you experience the relationship between your actions and the resulting action of the installation?

3) During the time of the experiment what did you experience as the result of using your voice?

4) What part of your voice if any at all were important in the installation?

5) During the time of the experiment to what extent if at all did your voice change?

6) Which sort of sounds had the most impact or were the most effective Could you make one of the sounds again now?

7) When you stopped making sounds what would happen to the forms?

Virtual construct:

8) During the time of the experiment what did you think the displayed form was?

9) During the time of the experiment what did you think your relationship to the form was?

10) During the time of the experiment would you say that you could control the form?

11) When you think about the experiment, how would you describe the shape of the form?

Time:

12) To what extent if at all did you experience a delay between your interaction and resulting form?

13) To what extent do you think that this influenced your interaction with the installation?

14) What do you think that this delay imposes on the relationship between you and the forms?

15) How did you experience the relationship between the time of the forms and the time of your location?

Embodiment:

16) To what extent if at all did you experience that the forms were linked to your presence?

17) Did you at anytime during the experiment experience that you could control the objects as easily as your body?

18) To what extent if at all did you identify with the forms?

Presence:

19) To what extent if at all did you feel engaged with the task of interacting?

20) To what extent if at all did you feel present in the environment?

21) During the time of the experiment where did you feel that you were located?

22) During the time of the experiment where would you say that the displayed forms were located?

23) To what extent if at all did you feel that there were times at which the environment became the dominant reality for you and you almost forgot about the surroundings of the lab?

24) If such moments did happen how would you describe where you were?

25) During the time of the experience which was the strongest on the whole, your experience of being in the environment or of being in the lab?

Space:

26) When you think back about your experiences do you remember it more as images you saw or somewhere you visited?

27) If you remember it as somewhere you visited how would you describe that place?

28) If you do remember it as somewhere you visited how would you describe where the forms were in relationship to you?

29) If you do remember it as somewhere you visited would you say that you present in the same space as the objects?

30) Where would you say that the space sited?

Space and interaction:

31) To what extent if at all did your interaction affect the space of the forms?

32) During the time of the experiment how would you describe the space of the forms during moments of silence?

33) When your were not using your voice where did your sense of presence place itself.

Multiuser questions:

34) During the experiment what was your relationship to the other user?

35) To what extent if at all did you experience that the other user was in the same place as you?

37) To what extent if at all did the other user have impact on the forms?

38) When you think back about your experience, do you remember this as more like just interacting with the forms or with another person?

39) To what extent of at all did you establish communication with the other user?

40) If you did establish communication at all how would you describe this communication?

41) To what extent were you and the other person in harmony during the course of the performance of the task?

Appendix B: User Manual

Log on to *Exodus* or, if you don't have permission, ask the friendly staff to assist you. Make sure the microphone is switched on and properly attached to the device (if in doubt ask staff). Open a UNIX shell and run the audio panel using the command `apanel`. Set the input device to `Line in`. Also, set the sample rate to 16kHz. Set the sensitivity of the input device to around 0, though not maximum. Check the box named `Meter` and test whether the display responds accordingly. If in doubt, check box named `Monitor` to get audio feedback from the local speakers (inside the CAVE), uncheck afterwards.

Next, use `cd` to change to the directory containing the program; if it is on a floppy disk, copy it to your home directory using `cp single ~`. If you only have source code at hand, you can copy it to a directory of your choice and compile it using the command `gmake`. This will result in the above file being generated. For this to work, it is important that the compilation instruction stated in the `GNUmakefile` is present in the same directory.

Now make sure the microphone is attached to your shirt, ideally near your collar. Unlock the CAVE processors from the menu by expanding the

toolchest and then selecting `Unlock Processors`. Start the program by typing its name (i.e. `single`) and hit `<Return>`.

Take off your shoes, take a pair of shutter glasses, open them and put them on. If you can't see through one glass, remove them, close them, and repeat the procedure, until this is rectified. Put on the head tracker making sure it is positioned correctly and supported by the glasses.

Now enter into the CAVE and *use your voice to manipulate the objects in an ordered way. Try to understand the relation between the action and response.* To end the program, go back to the terminal, move the mouse to the right until it becomes visible in the CAVE and press `<ESCAPE>`. If necessary, repeat pressing `<ESCAPE>` until program has shut down and the displays returned to the desktop environment.

Note: if the program crashes during a session, simply restart it after unlocking the processors. First, attempt to quit the old program in the way described above. If this fails, move the cursor back over the shell and press `<CTRL> + C`. Follow the above steps to restart the program.

The main steps are summarised in order below:

1.) Log on to Exodus.

2.) Switch on microphone and connect to device.

3.) Run audio panel using `apanel`.

   a. set input device to Line in;

   b. set sample rate to 16kHz;

   c. set sensitivity to around 0;

   d. verify that voice input is received by your local machine.

4.) Change to directory containing program

5.) Attach microphone to shirt.

6.) Unlock CAVE processors.

7.) Run program by typing `single`.

8.) prepare for CAVE and enter.

9.) Interact.

Appendix C: Planning and Execution

There were a number of factors that had to be considered during planning. First of all, I had to learn about graphics and virtual environments and in particular I needed to learn programming languages such as OpenGL and CAVELib in order to be able to program the CAVE™ environment. Then I needed to know about various aspects of speech processing and how to implement a number of related algorithms. The system and experiments needed to be designed, specified and implemented. All of these factors had to be arranged into an agenda. Finally, the data had to be evaluated and the thesis be written.

Given these tasks and the time of roughly six months, the following plan was developed. First, and most importantly, I had to be in a position to write applications for the CAVE™, so I started off learning OpenGL, while also reading background on VR. When I had a good enough command of it, I combined it with CAVELib and started writing applications for the CAVE™.

The next step was to start thinking about the system design. Once that was done, I could consult relevant work on speech processing and take from it what I needed for our implementation. Consequently, I started writing little programs that would perform a certain task. Writing speech processing software caused a lot more problems than expected, and this, in fact, imposed a delay on the experimentation, evaluation and write-up. This, in effect meant, that instead of having four weeks to carry out experiments in July, we were left with one week in the beginning of August. Although this was not an immediate danger with regards to combining the entire work into a thesis, but also constrained this issue further.

Regarding the schedule and actual sequence of events, they can be compared in the diagram below.

	Planned		Outcome	
March	-Learn OpenGL	-Read VR literature	-Learn OpenGL	-Read VR Literature
April	-Learn VR programming. Design system	-Read VR -Read speech processing	-Learn OpenGL Learn VR programming.	-Read VR
May	-Implement System	-speech processing	-Learn VR programming. -Design System	-Speech processing
June	-Implement System. -Testing. -Begin experiments.		-Implement System.	-Speech processing
July	-Experiments -Evaluation		Implement System.	
August	-Evaluation and write up dissertation		Single/multi-user experiments Evaluation and write up dissertation	

Appendix D: Contributions

In this work, we presented a new and cross-sensory way of interacting with virtual objects by employing intuitive characteristics of the human voice. We have visualised parameters of the human voice in a 3D immersive virtual environment and shown that users can gain control of the intuitive interface and learn to adapt to new and previously unseen tasks in virtual environments. We failed, however, to convincingly link successful control of an environment to a higher degree of presence, which, in our case, seemed to have been supported by the exceptional visual experience more than anything else. Nonetheless, we could associate a high degree of presence with a high degree of interaction, which seems to at least partially support the recent trend towards explaining presence through interaction (see §2.4 for details).

We gained a lot of insight about the impact of group behaviour on the actual experiments. This may have implications on the conduct of any future multi-user study, because it shows that there are matters that are currently not being controlled, measured or accounted for in the design of environments or conduct of standard experiments. Subjects are ultimately more important to the success of an experiment than these factors, so we need to find a way to incorporate more knowledge about human behaviour in social groups into our models that might help and overcome these problems.

Though an application such as this can not be used in computer-supported collaborative work (CSCW) environments cross-sensory interfaces bear great potential for use in arts or entertainment as almost every subject reported they had enjoyed the experiment and the environment a great deal.

References

[1] Ainsworth, W. A., Pitch change as a cue to syllabification, *Journal of Phonetics*, 14, pp. 257-264, 1986.

[2] Ainsworth, W. A., Lindsay, D., Identification and discrimination of Halliday's primary tones, *Proc. Inst. Acoustics*, 6, pp. 301-306, 1984.

[3] Bach-y-Rita, P., Sensory plasticity: Applications to a vision substitution system, *Acta Neurol. Scand. Vol. 43*, pp. 417-426, 1967.

[4] Bach-y-Rita, P., Kaczmarek, K.A., Mitchell, E.T., Garcia-Lara, J., Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note, *Journal of Rehabilitation R&D Vol. 35 (4)*, pp. 427-430, 1998.

[5] Benford, S., Greenhalgh, C., Lloyd, D., Crowded collaborative virtual environments , *Proceedings CHI1997*, 1997.

[6] Benford, S., Greenhalgh, C., Reynard, G., Brown, C., Koleva, B., Understanding and constructing shared spaces with mixed reality boundaries *ACM Transactions on Computer-Human Interaction*, *Vol. 5 (3),* pp182- 223, 1998.

[7] Benford, S., Bederson, B.B., Åkesson, K.-P., Bayon, V., Druin, A., Hansson, P., Hourcade, J.P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K.T., Stanton, D., Sundblad, Y., Taxén, G, Designing storytelling technologies to encourage collaboration between young children, *Human Factors in Computing Systems: CHI 2000 ACM Press*, 2000.

[8] Billinghurst, M., Kato, H., Collaborative Mixed Reality, *Mixed Reality – Merging Real and Virtual Worlds*, pp 261-284; Springer Verlag, Berlin, 1999.

[9] Bliss, J.P., Tidwell, P.D., Guest, M.A., The effectiveness of virtual reality for administering spatial navigation training to fire fighters, *Presence Vol. 6*, pp73-86, 1997.

[10] Brooks, B.M., Attree, E.A., Rose, F.D., Clifford, B.R., Leadbetter A.G., The specificity of memory enhancement during interaction with a virtual environment, *Memory Vol. 7 (1)*, pp 65-78, 1999.

[11] Cooke, M., Beet, S., Crawford, M. (editors), Visual representation of speech signals, John Wiley & Sons, 1993.

[12] Cooley, J. W., Tukey, J. W., An algorithm for the machine calculation of complex Fourier series, *Math. Comput.* 19, p.297, 1965.

[13] Degenaar, M., Molyneux's Problem: three centuries of discussion on the perception of forms, Kluwer Academic Publishers, 1996.

[14] Durlach, N., Allen, G., Darken, R., Garnett, R.L., Loomis, J., Templeman, J., von Wiegand, T., E. Virtual Environments and the enhancement of spatial behaviour: towards a comprehensive research agenda, *Presence Vol. 9 (6)*, pp. 593-615, 2000.

[15] Ellis, S. R., Pictorial communications in virtual and real environments, Taylor and Francis, London, 1993.

[16] Fahlén, L. E., Brown, C. G., Ståhl, O., Carlsson, C., Space Based Model for User Interaction in Shared Synthetic Environments, *Interchi '93*, 1993.

[17] Fant G., Acoustic Theory of Speech Production, Mounton, The Hagues, 1960.

[18] Frigo, M., Johnson, S.G., The Fastest Fourier Transform in the West, MIT-LCS-TR-728, 1997.

[19] Gibson, J.J., The ecological approach to visual perception, Houghton Mifflin, Boston, 1979.

[20] Glaser, B., Strauss A., The discovery of grounded theory: strategies for qualitative research, Aldine Publishing Company, Chicago, 1967.

[21] Glenberg, A.M, What memory is for. *Behavioral and Brain Sciences, 20,* 1-19, 1997.

[22] Gold, B., Rabiner, L., Parallel processing techniques for estimating pitch periods in the time domain. *Journal of the Acoustical Society of America*, Vol. 46, pp. 442-448, 1969.

[23] Gruenz, O., Scott, L. O., Extraction and portrayal of pitch of speech sounds, *Journal of the Acoustical Society of America*, Vol. 21, pp. 487-495, 1949.

[24] Heidegger, M., Sein und Zeit, M. Niemeyer, Tbg., 1993 (original: 1927).

[25] Held, R.M., Durlach, N.I., Telepresence, *Presence Vol. 1 (2)*, pp. 109-112, 1992.

[26] Hermes, D.J, Pitch Analysis, in Cooke, M., Beet, S., Crawford, M. (editors), Visual representation of speech signals, pp. 3-25, John Wiley & Sons, 1993.

[27] Hess, W., Pitch determination of speech signals, Springer Verlag, Berlin, 1983.

[28] Hoberman, P. Parés, N, Parés, R., El Ball del Fanalet or Lightpools, *Proceedings of International Conference on Virtual Systems and Multimedia '99*, pp. 270-276, 1999.

[29] Ingram, R., Benford, S., The application of legibility techniques to enhance information visualisations, *Computer Journal Vol. 39 (10),* 1996.

[30] Ishii, H., Kobayashi, M., Arita, K., Iterative design of seamless collaboration Media. *Communications of the ACM*, Vol. 37, No. 8, pp. 83-97, 1994.

[31] Ishii, H, Ullmer, B., Tangible Bits: towards seamless interfaces between people, bits and atoms, *CHI97*, 1997.

[32] Klatt, D., H., Software for a cascade/parallel formant synthesizer, *Journal of the Acoustical Society of America*, 82, pp. 737-793, 1980.

[33] Klatt, D. H., Klatt, L. C., Analysis, synthesis and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America*, 87, pp. 820-857, 1990.

[34] Kolb, B, Brain Plasticity and behaviour, Lawrence Erlbaum, NJ, 1995.
Koleva B., Schnädelbach H., Benford S., Greenhalgh C. Traversable interfaces between real and virtual worlds, *Proceedings CHI2000*, 2000.

[35] Meijer, P.B.L., An Experimental System for Auditory Image Representations, *IEEE Transactions on Biomedical Engineering*, *Vol. 39. (2)*, pp. 112-121, 1992.

[36] Milgram, P., Kishino, F., A taxonomy of mixed reality visual displays, IEICE Transactions on Information Systems, Vol E77-D, NO.12, 1994.

[37] Minsky, M., Telepresence*. Omni*, pp. 45-51, 1980.

[38] Morgan, M.J., Molyneux's Question - Vision, Touch and the Philosophy of Perception. Cambridge University Press, 1977.

[39] Naimark, M., Elements of realspace imaging: a proposed taxonomy, *Proceedings of SPIE 1457*, 1991.

[40] Naimark, M., Elements of realspace imaging, *Technical Report, Apple Multimedia Lab*, 1992.

[41] Noll A. M., (1967) Cepstrum pitch determination, J *Journal of the Acoustical Society of America*, 1967, Vol. 41, pp. 293-309.

[42] Oren, T., Desigining a new medium, in Laurel, B. (ed.), *The art of human-computer interface design*, Addison-Wesley, Reading, MA, pp. 467-479, 1990.

[43] Ramshaw, L., Blossoming: a connect the dots approach to splines, Technical Report 19, Digital Systems Research Center, 130 Lytton Avenue, Palo Alto, California 94301, 1987.

[44] Recanzone, G.H., Merzenich, M.M., Jenkins, W.M., Frequency discrimination training engaging a restricted skin surface results in an emergence of a cutaneous response zone in the cortical area, *Journal of Neurophisology 67*, pp 1057-1070, 1992.

[45] Robinett, W., Synthetic Experience: a proposed taxonomy, *Presence Vol. 1 (2)*, pp 229-247, 1992.

[46] Rose, F.D., Attree, E.A., Brooks, B.M., Andrews, T.K., Learning and memory in virtual environments: a role in neurorehabilitation? Questions (and occasional answers) from the University of East London, *Presence Vol. 10 (4)*, pp. 345-358, 2001.

[47] Rosen, S., Howell, Signals and systems for speech and hearing, Academic Press, London, 1991.

[48] Schmalstieg, D., Fuhrmann, A., Szalavari, Z., Gervautz, M., Studierstube – an environment for collaboration in augmented reality, *CVE '96 Workshop Proceedings*, 1996.

[49] Schubert, T., and Regenbrecht, H., embodied presence in virtual environments, *in Paton, R., Neilson, I. (Eds.), Visual Representations and Interpretations*, 1999.

[50] Schuemie, M. J., van der Mast, C.A.P.G., Presence: interacting in VR?, *Proceedings of the TWLT 15*, pp. 213-217, 1999.

[51] Schroeder, M. R., Period histogram and product spectrum: new methods for fundamental frequency measurement, *Journal of the Acoustical Society of America*, 1968, Vol. 43, pp. 829-834.

[52] Seidel, R. J., Chatelier, P. R. (eds.), Virtual Reality. Training's Future? Perspectives on Virtual Reality and Related Emerging Technologies, Plenum Press, NY, 1997.

[53] Sheridan, T.B., Musings on telepresence and virtual presence, *Presence Vol. 1 (1)*, 1992.

[54] Shneiderman, B., Designing the user interface, Addison-Wesley, Reading, MA, 1992.

[55] Slater, M., Usoh, M., Body-centred interaction in immersive virtual environments, in Magnenat-Thalmann, N., Thalmann D. (eds.) *Artificial Life and Virtual Reality*, pp. 125 -148. John Wiley and Sons, 1994a.

[56] Slater, M., Usoh, M., Steed, A. (1994), Steps and Ladders in Virtual Reality, *ACM Proceedings of VRST '94*, pp. 45-54, 1994b.

[57] Slater, M., Usoh, M., Steed, A., Depth of Presence in Virtual Environments, *Presence Vol. 3 (2*), pp. 130-144, 1994c.

[58] Slater, M., Sagadic, A., Usoh, M., Schroeder, R., Small Group Behaviour in a Virtual and Real Environment: A Comparative Study, *BT Presence Workshop, Martlesham Heath*, 1998.

[59] Slater, M., Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire, *Presence, Vol. 8 (5)*, pp 560-565, 1999.

[60] Slater, M. and Steed, A., A Virtual Presence Counter, *Presence Vo. 9 (5),* pp. 413-434, 2000.

[61] Stanney, K. M., Mourant, R.R., Kennedy, R.S., Human Factors Issues in Virtual Environments: A review of the literature, *Presence Vol. 7 (4)*, pp. 327-351, 1998.

[62] Stanton, D, Wilson, P., Foreman, N., Duffy, H., Virtual environments as spatial training aids for children and adults with physical disabilities, *Proceedings of International Conference on Disability, Virtual Reality, and Associated technologies 2000*, pp. 123-128, 2000.

[63] Strauss, A., Corbin, J., Basics of qualitative research, Sage, Newbury Park, 1990.

[64] Steuer, J., Defining Virtual Reality: Dimensions Determining Telepresence, *Journal of Communication , 42 (4),* pp. 72-93, 1992.

[65] Synnot, A., The body social: symbolism, self, society, Routledge, New York, 1993.

[66] Witmer, B.G., Singer, M.J., Measuring Presence in Virtual Environments: A Presence Questionnaire, *Presence Vol. 7 (3)*, pp. 225-240, 1998.

[67] Witten, I., Principles of Computer Speech, Academic Press, London, 1982.

[68] Zagal, J. P., Nussbaum, M., Rosas, R., A model to support the design of multiplayer games, *Presence Vol. 9 (5)*, pp 448-462, 2000.

[69] Zahorik, P., Jenison, R.L., Presence as Being-in-the-world, *Presence Vol. 7 (1)*, pp. 78-89, 1998.

[70] Zeltzer, D., Autonomy, interaction presence, *Presence Vol. 1 (1)*, p127-132, 1994.