UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

Increasing presence in first-person virtual environments through auditory intefaces: an analytical approach to adaptative sound and music

Incrementar la presencia en entornos virtuales en primera persona a través de interfaces auditivas: un acercamiento analítico al sonido y la música adaptativos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Manuel López Ibáñez

Directores

Pablo Gervás Gómez-Navarro
Federico Peinado Gil

Madrid

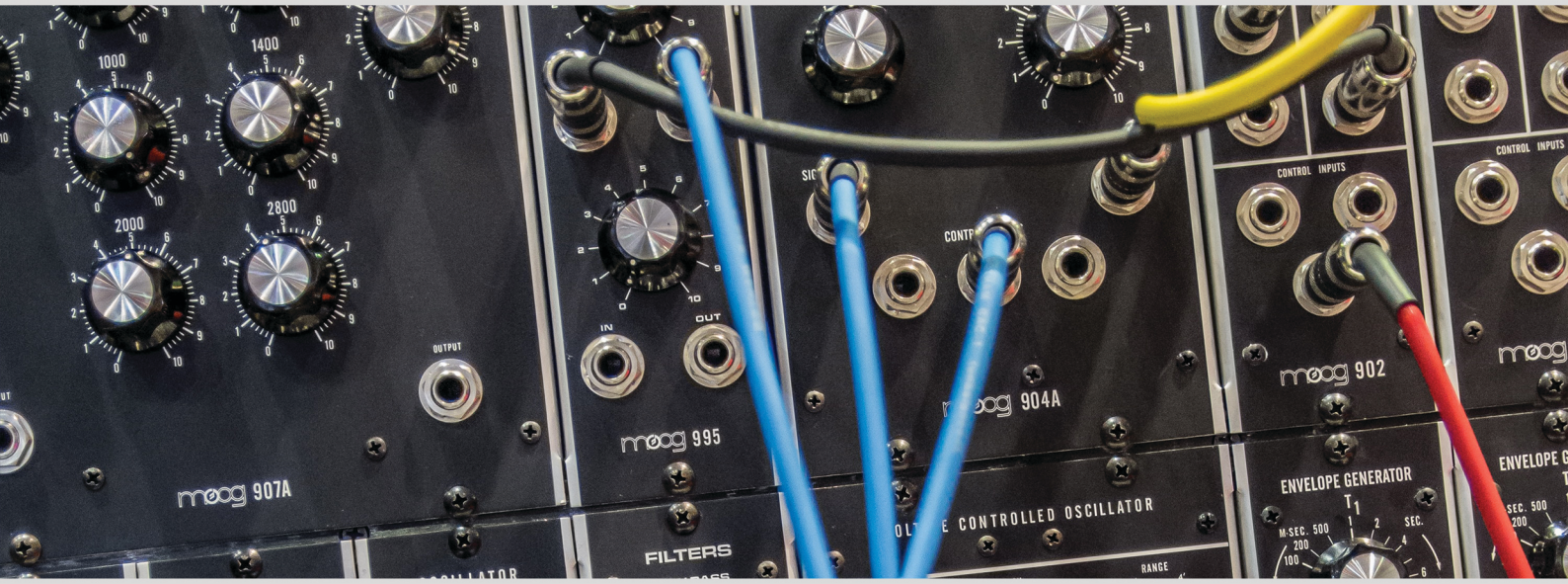
Increasing presence in first-person virtual environments through auditory interfaces

An analytical approach to adaptive sound and music

Incrementar la presencia en entornos virtuales en primera persona a través de interfaces auditivas

Un acercamiento analítico al sonido y la música adaptativos

Directores:
Pablo Gervás Gómez-Navarro
Federico Peinado Gil



MANUEL LÓPEZ IBÁÑEZ

Facultad de Informática
Universidad Complutense de Madrid

Increasing presence in first-person virtual environments through auditory interfaces

An analytical approach to adaptive sound and music

Incrementar la presencia en entornos virtuales en primera persona a través de interfaces auditivas

Un acercamiento analítico al sonido y la música adaptativos

Manuel López Ibáñez

Directores:
Pablo Gervás Gómez-Navarro, Federico Peinado Gil

Universidad Complutense de Madrid

13 de septiembre de 2019

UNIVERSIDAD COMPLUTENSE MADRID

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña. Manuel López Ibáñez,
estudiante en el Programa de Doctorado en Ingeniería Informática, RD99/2011,
de la Facultad de Informática de la Universidad Complutense de Madrid, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Increasing presence in first-person virtual environments through auditory interfaces: An analytical approach to adaptive sound and music

Incrementar la presencia en entornos virtuales en primera persona a través de interfaces auditivas: Un acercamiento analítico al sonido y la música adaptativos

y dirigida por: Pablo Gervás Gómez-Navarro y Federico Peinado Gil

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita.

Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada de conformidad con el ordenamiento jurídico vigente.

En Madrid, a 3 de octubre de 2019

Fdo.: MANUEL LÓPEZ IBÁÑEZ

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Dedication and acknowledgements

To Lucía, who made me listen.

This research has been supported by Banco Santander, in cooperation with Fundación UCM, in the form of a predoctoral scholarship (CT2716 - CT2816).

Additional funding was provided by project NarraKit VR: Interfaces de Comunicación Narrativa para Aplicaciones de Realidad Virtual (PR41/17-21016) (Banco Santander-UCM).

Support was also provided through EC funding for the project WHIM 611560 by FP7, the ICT theme, and the Future Emerging Technologies (FET) programme.

TABLE OF CONTENTS

List of Tables

List of Figures

Abstract

The popularisation of virtual reality devices has brought with it an increased need of telepresence and player immersion in video games. This goals are often pursued through more realistic computer graphics and sound; however, invasive graphical user interfaces are still present in industry standard products for VR, even though previous research has advised against them in order to reach better results in immersion. Non-visual, multimodal communication channels are explored throughout this thesis as a means of reducing the amount of graphical elements needed in head-up displays while increasing telepresence. Thus, the main goals of this research are to find the optimal channels that allow for semantic communication without recurring to visual interfaces, while reducing the general number of extra-diegetic elements in a video game, and to develop a total of six software applications in order to validate the obtained knowledge in real-life scenarios. The central piece of software produced as a result of this process is called LitSens, and consists of an adaptive music generator which takes human emotions as inputs.

Prior to the description of each research process, a review of currently available literature in several relevant fields is included. A unified definition of the concepts of "presence" and "immersion" is given, and the central themes of this thesis are presented in detail.

Later on, an experiment which connects virtual movement, presence and simulator sickness is described, and the conclusion that virtual, continuous locomotion is possible in virtual reality under certain circumstances is reached. Next, a brief experiment on how sound can influence player behavioural tendencies during *gameplay* is included. After realising that auditory interfaces have enough potential for replacing visual interfaces, new sound spatialisation techniques are developed and tested in a playable video game prototype.

LitSens starts its development at this point, as a system that can produce adaptive music in order to increase player presence. The resulting software architecture is tested in two popular video game engines: Unreal Engine and Unity. Additional experimentation and validation is conducted to check if LitSens can influence player behaviour and feelings in real time, with positive results. Lastly, a head gesture recognition system for VR devices based on neural networks is added, to allow for immediate player feedback while running this software.

The main contributions that can be extracted from this research are: the building of the adaptive music generator named LitSens, the discovery of the relationship between non-accelerated movement, presence and simulator sickness, the creation of new methods for player orientation in 3D environments (based on auditory spatialisation) and the building of a neural network with the ability to recognise emotions in players by only collecting data from head movements.

Keywords: `virtual reality, video games, telepresence, music, sound design, virtual environments`

Resumen

La popularización de los dispositivos de realidad virtual ha traído consigo una mayor necesidad de presencia e inmersión para los jugadores de videojuegos. Habitualmente se intenta cumplir con dicha necesidad a través de gráficos y sonido por ordenador más realistas; no obstante, las interfaces gráficas de usuario muy invasivas aún son un estándar en la industria del videojuego de RV, incluso si se tiene en cuenta que varias investigaciones previas a la redacción de este texto recomiendan no utilizarlas para conseguir un resultado más inmersivo. A lo largo de esta tesis, varios canales de comunicación multimodales y no visuales son explorados con el fin de reducir la cantidad de elementos gráficos extradiegéticos necesarios en las capas de las interfaces gráficas de usuario destinadas a la representación de datos, todo ello mientras se logra un aumento de la sensación de presencia. Por tanto, los principales objetivos de esta investigación son encontrar los canales óptimos para efectuar comunicación semántica sin recurrir a interfaces visuales —a la vez que se reduce el número de elementos extradiegéticos en un videojuego— y desarrollar un total de seis aplicaciones con el objetivo de validar todo el conocimiento obtenido mediante prototipos similares a videojuegos comerciales. De todos ellos, el más importante es LitSens: un generador de música adaptativa que toma como entradas emociones humanas.

Antes de pasar a describir en detalle cada uno de los procesos de experimentación que dan forma a esta tesis, se incluye una revisión de la literatura académica disponible actualmente acerca de una serie de asuntos relevantes en este contexto. Se proporciona, asimismo, una definición unificada de los conceptos de «presencia» e «inmersión» y se describen en detalle los temas centrales de esta investigación.

Acto seguido, se presentan los resultados obtenidos a través de un experimento que conecta el movimiento virtual, la presencia y la simulator sickness (o "enfermedad del simulador"), y que permite alcanzar la conclusión de que el desplazamiento simulado y continuo es posible en entornos virtuales que cumplen con una serie de características. A continuación, se describe un breve experimento acerca de cómo el sonido puede influir en el comportamiento de los jugadores. Tras comprobar que las interfaces auditivas tienen suficiente potencial como para reemplazar a las interfaces visuales, se presentan una serie de nuevas técnicas de espacialización sonora, que son validadas a través de un prototipo jugable.

LitSens comienza a desarrollarse en esta fase de la investigación y consiste, en un principio, en un sistema para la generación de música adaptativa con el fin de incrementar la presencia. La arquitectura de software resultante es puesta a prueba en dos motores de videojuegos muy populares: Unreal Engine y Unity. Posteriormente, se realizan nuevos experimentos con la intención de validar el funcionamiento de LitSens y comprobar si realmente puede influir sobre el comportamiento y las emociones de los usuarios en tiempo real; los resultados obtenidos al respecto son positivos. Por último, se añade a esta aplicación un sistema de reconocimiento de emociones en función de los gestos de la cabeza mientras se lleva puesto un

dispositivo de realidad virtual, basado en redes neuronales, que permite retroalimentación inmediata por parte de los usuarios mientras está en ejecución.

Las principales contribuciones que pueden ser extraídas de este proceso investigador son: la construcción de un generador de música adaptativa (LitSens), el descubrimiento de la relación existente entre el movimiento no acelerado, la presencia y la simulator sickness, la creación de nuevos métodos que permiten la orientación de los jugadores en entornos tridimensionales (basada en espacialización auditiva) y la construcción de una red neuronal con la capacidad de reconocer emociones en jugadores tan solo a partir de movimientos de la cabeza.

Palabras clave: realidad virtual, videojuegos, presencia, música, diseño sonoro, entornos virtuales

CHAPTER 1

INTRODUCTION

"Aion is a child at play, playing draughts; the kingship is a child's."

Heraclitus

We live in a world full of virtual experiences, and computer software has taken a decisive role in the process of creation of modern narratives. Nowadays, video games constitute one of the most popular frameworks for interactive storytelling, in terms of both business value ($137.9 billion in 2018) and audience (2.3 billion active users during the same year) [118].

Most current video games are based on virtual environments, which for this context can be defined as "interactive, head-referenced computer displays that give users the illusion of displacement to another location" [26].

The visual dominance of contemporary Western culture [80] —in addition to the big influence of cinema, television and the Internet over game designers and developers— has determined the usual characteristics of virtual environments; they tend to allow for two-dimensional (2D) or three-dimensional (3D) representation of images and text, and in many cases try to achieve player immersion, either through the technology itself [43], narrative techniques [72] or a combination of both. Computer graphics play, therefore, a very important role in this process, and have improved greatly since the first video games were published.

With the advent of commercial virtual reality (VR from now on), the need for better

immersion and presence [103] has increased, as will be explained in detail in later sections. Better or more realistic graphics are not the only (and not necessarily the better) way to maintain an acceptable level of presence in VR, and many important developers in this industry have shifted focus towards audio, which had remained in a discrete position since the popularization of surround sound for video games. Such is the case of Valve, with Steam Audio[1] and Google, with Resonance Audio[2]. The reason for this trend is that sound highly contributes to the achievement of presence in first-person virtual experiences, as several recent works have shown [51, 75].

The intention behind this Ph.D. thesis is to contribute to the fields of game audio and presence analysis, while maintaining an open, multidisciplinary approach. More specifically, the aim is to utilise auditory stimuli not only as a reliable representation of how certain environments would sound in the real world, but as a tool for game creators and an alternative to all-pervading graphical user interfaces in VR. In this process, the need for more presence and immersion is always taken into account, along with the industry-wide wish for better locomotion and player orientation systems in VR. Technical advancements in the domain of Computer Science are presented, but these results, due to sheer necessity, are often intertwined and interact with other fields, such as Psychology and Creative Arts.

During this whole research, experimentation with real subjects was a priority. For this very reason, six playable video game prototypes were built in order to create environments representing the challenges and problems that needed to be solved through the course of this thesis. 5 of these software applications (see Appendix E) were designed for testing and academic purposes only, while one is a commercial video game that makes use of some of the advancements presented here.

The most recent research efforts in this process were dedicated to the development of LitSens: a software application capable of generating adaptive sound, making use of many of the conclusions that can be extracted from this work. More than a unique application, LitSens aims to be a general architecture for any video game engine, and has been currently implemented in two of the most common commercial engines available, as will be explained later.

1.1 Motivation

The main goal of this research was to find a way to increase presence, as is defined by Sheridan [103], Barfield and Weghorst [5], and Mestre, Fuchs, Berthoz and Vercher [77], in the context of virtual narrative experiences in first-person perspective which are controllable

[1] `https://valvesoftware.github.io/steam-audio/`
[2] `https://developers.google.com/resonance-audio/`

through specific hardware (HMDs, different types of game controllers, a mouse, a keyboard or any combination of them). The contents of this thesis revolve around video games, as they represent one of the most complete multimodal interactive experiences, with a variety of data inputs and outputs [12].

Specifically, the following tasks were considered before starting this study:

- To find the optimal communication channels with potential players to produce an experience which increments general presence through resources available in regular video games (user interface, audio, control hardware, etc.).
- To reduce the number of extra-diegetic elements in graphical user interfaces, and to validate their effects on presence.
- To evaluate the efficiency of different linguistic communication methods in a first-person video game, taking into account the need to maintain or increase presence.
- To broaden the scope of multimodality in video games, in order to explore new communication channels which could be of use when trying to increase presence. One way to do this would be to add specific semantic information to channels that usually do not convey it, such as audio.
- To design and implement a total of six software applications that validate the knowledge obtained through this research.

These were later modified and adapted in accordance with the advancements and findings that came with the research process itself, described in Chapters 3 to 8, and resulted in the final contributions listed in Chapter 9.

1.2 Research plan

This research went through 7 different phases, and a variety of research methodologies were employed to reach the conclusions detailed in Chapter 9. The following outline briefly explains each of them.

- **Definition of a frame of reference:** During this phase, previous work was analysed in order to better specify the scope of the research. The main hypotheses were also formulated at this point.
- **Classification of communication channels:** After reaching the conclusion that there would be a need to test user reactions to certain strategies, several multimodal

communication channels were established as candidates to be analysed during experimentation. Those channels were, mainly: written or oral text, perceived time (from the perspective of Schneider, Kisby and Flint [101]), touch and gestures [11] and audio.

- **Building of playable prototypes:** At this stage, several playable prototypes started development. These were meant to experiment with real video game users, and revolved around the concept of presence, as will be detailed in the following Chapters. Several existent technologies were utilised during this phase, such as Unreal Engine 4, Unity, OpenVR or Scikit-learn. Their use will also be explained during each relevant section of this thesis. The conjunction of all these small experiences gave room to the development of a bigger software application that was called LitSens, and which made use of the most relevant findings of this research.
- **Results analysis and method extraction:** After experimenting with each of the aforementioned prototypes, retrieved information was analysed. Some preliminary guidelines and methods, like the ones at the ending of Chapter 4, were extracted and applied to the consecutive iterations of LitSens.
- **Consolidation or rebuttal of the main hypotheses:** After the last steps, there was enough information available to adapt the theoretical framework and the main hypotheses to the new findings.
- **Further improvement of LitSens:** Considering the new framework established by previous experimentation, LitSens was developed until it reached its current shape, and used to experiment further.
- **Publishing of results:** During this whole process, six academic articles were written, containing the main advancements in this research. Along with them, the contents of this Ph. D. thesis were also produced.

1.3 Hypotheses

As was described in the previous section, a series of hypotheses were formulated at the beginning of the research process. These were the starting point for the rest of the advancements made, and provided a foundation for the final contributions, described during Chapter 9. They also represent the evolution of the reasoning behind this Ph. D. thesis.

- Player-controlled cinematic movement is possible in VR without producing simulator sickness or decreasing presence, as long as certain parameters are supervised and tweaked.

- Traditional graphical user interfaces are not effective when orienting or aiding player locomotion in virtual reality.
- Sound and music can influence how a player behaves and moves in a virtual environment, as long as several key parameters are changed.
- Sound and music can be used as a proxy for graphical user interfaces in player orientation scenarios.
- Player emotions and psychological profiles can have an influence over how they are affected by sound and music when orienting themselves in a virtual environment.
- A low-latency spatialisation technique, based on the application of low pass filters to certain in-game sounds, can provide better results in terms of player orientation and position identification in virtual environments than fully-fledged 3D audio.
- Adaptive music can be helpful when trying to provide real time feedback to players in a virtual environment, as well as increasing presence.
- Player behaviour can be influenced through an adaptive music system.
- Player gestures can be used as an input for an adaptive music system based on emotion recognition.

These preliminary ideas will be described and contextualised in more detail in each relevant chapter.

1.4 Contributions

The prospective contributions of this thesis can be classified as *primary* or *secondary*, depending on their importance for the described fields of research.

The most important contributions are the following:

- To discover new methods for improved player orientation in 3D environments, with a focus on auditory interfaces, and aiming to reduce the amount of graphical elements in a GUI for VR.
- To build a multiplatform software application that can play adaptive music depending on emotional inputs received from a video game session.
- To establish a noninvasive method for knowing the essential emotional state of users while they play, based on gestures.

And the secondary contributions are:

- To evaluate the advantages and disadvantages of the most popular methods for creating adaptive and dynamic music, according to currently available academic literature.
- To find a method for controlling simulator sickness while increasing presence through virtual locomotion in VR environments.

1.5 Resources and equipment

Material and software resources needed during the development of this research were limited to the ones present in the following list:

- Video game engines: Unreal Engine 4 and Unity 3D. They were used for the purpose of developing the different pieces of software that will be described later.
- Computer: A PC meeting the requirements needed to launch virtual reality games. It was utilised for both developing and experimentation.
- HMDs: Oculus Rift and HTC Vive, two of the most popular headsets for virtual reality. They worked as devices by which the developed video games could be experienced, but also as trackers for several postural indicators.
- Game controllers: Xbox One controller, Oculus Touch and Vive Controllers, used to control the already-mentioned video games.
- Physical spaces: Labs and classrooms available for experimentation with users.

1.6 Structure

This document is organised in a total of 9 Chapters, 6 appendices and a bibliography. The first 2 Chapters, which have an introductory nature, are this foreword and an analysis of the reviewed literature. Chapters 3 to 8 constitute the most important part of this document, as they explain the different research processes and findings that shaped this thesis (by order of appearance): the relationship between movement, presence and simulator sickness, the influence that sound can have over gameplay, new techniques for sound spatialisation, the different iterations in the development of the core software for this Ph. D. thesis and the description of a gestural recogniser based on neural networks.

Chapter 9 discloses the main conclusions that can be drawn from the results presented earlier, after discussing the reliability of all the obtained data. The appendices contain data that is essential to properly understand the contents of this thesis: questionnaires used during experimentation, experiment designs and diagrams, a list of all software developed in the context of this research and a list of all the academic articles that support the main assertions made here. Lastly, the bibliography is organised by order of appearance in the main text, and is placed at the end of the document.

CHAPTER 2

PREVIOUS WORK

"Life can only be understood backwards, but it must be lived forwards."

Søren Kierkegaard

In this section, a series of concepts are introduced in order to provide a frame of reference for the research described in the next chapters. Firstly, an explanation on the notions of presence and immersion is included, as they constitute the fundamental concepts upon which this thesis is based. Later on, the subject of emotional audio is introduced, in connection with the existence of different psychological profiles in humans, which can influence how sound is perceived —from an emotional perspective— in certain situations.

A self-assessment method is also described, useful when trying to understand how users perceive emotional sound, while maintaining an agile experimentation process due to its simplicity and ease of use for subjects.

Besides, technical details on 3D sound spatialisation in virtual worlds are also given, in order to better understand the approaches taken throughout this thesis, and the reasons why they were chosen. As LitSens, the main piece of software developed during this research process, is an adaptive music system, details on the current literature on dynamic music are also included.

After this, a brief reflection on emotion extraction is also contained in this chapter, so as

to provide a better understanding of how LitSens evolved from reading textual emotions to automatically identifying certain basic feelings in gestural patterns. These basic emotions are also related to player behavioural tendencies, and the state of current academic literature is analysed in order to provide a solid foundation for Chapter 7.

Lastly, the subject of artificial intelligence in relationship with music generation is addressed, as well as the reasons behind the technical choices made during experimentation in Chapter 8, where two types of neural networks were used for gesture recognition and live music adaptation.

2.1 Presence, immersion and simulator sickness

Presence [5, 77, 103] is a fundamental concept for understanding the motivations behind this thesis. It can be defined as a general measure of the capacity of an interactive software application (usually involving virtual environments) to make a user experience a feeling of "being there"; "there" being the virtual environment in question.

However, this concept can be broken down into several, more specific ideas [9, 64]. The most common ones in the reviewed literature are:

- Telepresence: Measures the capacity of a given virtual world to create a *perceptual* illusion, in a strictly sensory way. Telepresence depends on how strong is the feeling of physical teleportation to the virtual environment.

- Social presence: Estimates the feeling of being in a virtual world with another person. Only social interactions themselves are taken into account in this case, as a subject could perfectly be having a realistic social exchange in an unrealistic virtual world.

- Presence as "mental transportation": Measures the quality of the transportation of a user's conscience into a virtual environment, and does not take into account the physical sense of teleportation. This last assertion is relevant, as a given subject could experience mental transportation without having physical feelings of "being there" —such is the case in an engaging first-person video game.

Throughout this work, the concept of social presence will be dismissed, as no social interactions took place during research and experimentation.

Immersion, on the other hand, stands for, according to the definition given by Mestre [77], "what the technology delivers from an objective point of view." Immersion is not a perceptual measure, and refers to the resemblance of VR devices to the senses and organs humans use to perceive the world around them. Also, as Slater and Wilbur put it, immersion is "a

description of a technology, and describes the extent to which the computer displays are capable of delivering an inclusive, extensive, surrounding and vivid illusion of reality to the senses of a human participant" [105].

Very related to presence is the concept of simulator sickness (SS from now on) [21, 50, 96]. It is described in the reviewed literature as a feeling of unease that arises from immersive virtual movement. Oculomotor discomfort, one of the diagnostic subscales of the Simulator Sickness Questionnaire [50], increases with virtual walking or flying, according to early research carried out by Usoh, Arthur, Whitton et. al. [115], and more recent studies also associate SS with first-person video games [76] and VR experiences [102].

SS must not be confused with motion sickness (MS from now on). Whereas MS occurs when there is physical movement in an environment that does not appear to be moving from the user's perspective (e. g.: a car), SS happens when there is no physical movement in an environment which does seem to move (e. g.: a VR video game).

Besides, SS is inversely proportional to presence [115], and it must be reduced as much as possible to obtain a pleasant user experience. The relationship between SS and movement in VR, however, has been questioned by authors like Serge and Moss [102], who state that SS did not change when switching between observational and navigational tasks while using an Oculus Rift, one of the most popular head-mounted displays (henceforth, HMDs). In any case, SS did rise over time in both types of experience.

The most obvious method for reducing SS is to avoid movement as much as possible during a virtual experience. However, more recent research shows that, while using modern HMDs, SS can be controlled by methods which do not imply removing movement at all. Here are some of these options:

- Sun, Patney, Wei et. al. [110] propose a system which uses eye tracking to find moments of temporary blindness during rapid eye movement (saccadic supression) to redirect players while walking; the redirection itself happens exclusively during these moments of blindness, only affects the in-game camera, and thus is unnoticeable by users. This could allow for virtually walking while maintaining a stationary position, and would drastically reduce SS accumulation when moving.

- Fernandes and Feiner [29], on the other hand, dynamically reduce field of view (FOV) taking into account movement and rotation speed, and achieve reduced levels of SS. This method has already been used in commercial VR video games such as "Eagle Flight"[1]. The reason for its success is that it takes advantage of monocular field of vision (see Figure 2.1): it keeps the binocular field focused at all times, while obscuring the monocular area. This guarantees clear vision in the center of the FOV at all times.

[1] `https://www.ubisoft.com/en-gb/game/eagle-flight/`

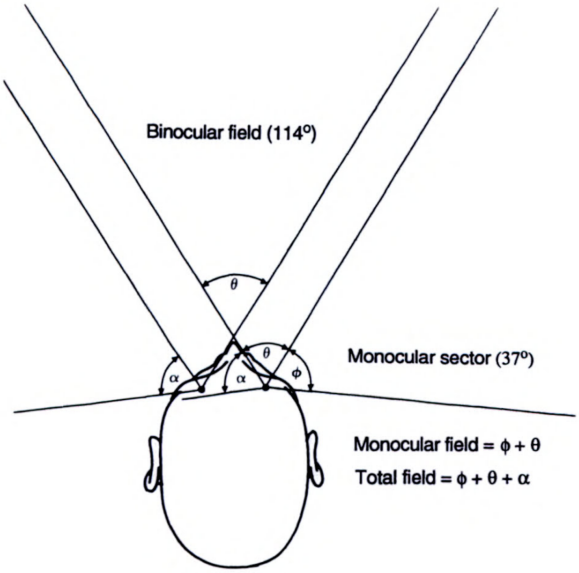


FIGURE 2.1. Representation of the human field of vision. Figure reproduced from [45].

– Another common option for reducing SS while in VR or first-person experiences is to let the player move only at a constant speed, with as few accelerations as possible. This means, for example, that a ramp works better than stairs when moving players vertically [20], because it allows for a smoother climbing movement.

– Some experts in this field, like Jason Jerald [47], also suggest that reducing texture complexity can help when trying to avoid SS in virtual environments. This means avoiding certain kinds of patterns, such as detailed meshes or textiles, gratings, etc.

– Though less practical, the utilisation of additional hardware peripherals is also a possibility. Devices like the Virtuix Omni[2] remove the problem of SS by actually letting players walk and move while they experience a VR video game, by means of a 360-degree treadmill. However, this sort of solution requires a big and expensive device, and is not oriented to the general public.

– Lastly, the most used technique nowadays consists in simply letting players *teleport* themselves to different locations, usually through a system like the one depicted in Figure 2.2, from the game "Robo Recall"[3].

[2] https://www.virtuix.com/
[3] https://www.epicgames.com/roborecall/en-US/home

FIGURE 2.2. Screenshot from the VR game *Robo Recall*, in which a teleportation system is depicted.

2.2 Sound, emotions and psychological profiles

Audio is another important factor when it comes to analysing the fluctuations of presence in a virtual experience, as these concepts are both strictly related to emotions. There exists a broad academic consensus when it comes to analysing the relevance of musical and auditory stimuli in the generation of emotions in humans [121, 127], and sound is also known to be able to produce certain conducts in listeners, as is laid out by Kivland [52]. Besides, there is a strong relationship between hearing and vision, and some shapes and colours [120] are often linked to certain types of sound.

In 1996, Richard Bartle [6] devised a psychological test which, taking into account the dominant emotions and attitudinal tendencies of players in an online video game, classified them in four big groups:

1. Socialiser: This group consists of people who like to share and dialogue in a video game. They usually like to empathise with the characters in the virtual world, and enjoy being told stories about human desires and emotions.

2. Killer: Killers tend to exert complete domination over the gaming environment. They like competition and ruling over the virtual world, and they are usually more open to making great efforts to achieve victory.

3. Achiever: Achievers like to show off and boast about their fictional characters. They try to complete and accumulate as much as possible, and enjoy a challenge as long as victory is reflected in any "palpable" way (like an achievement badge, a trinket, etc.). *Speedrunners* are frequently achievers who try to find new challenges outside of the original gaming experience. It is not uncommon for this kind of player to upload their feats to the internet in the shape of videos or screenshots.

4. Explorer: Explorers try to discover and comprehend all the little details present in the virtual world. They enjoy finding new and interesting things to do, and tend to memorise the layout of in-game levels and maps. They value knowledge and experimentation, and frequently like open worlds and sandbox experiences.

2.3 The Self-Assessment Manikin Test

During this research, a test that allowed for self-evaluation of an emotional state was needed, as a big part of this thesis has its foundations on improving presence by adapting audio to emotional responses. Specifically, the Self-Assessment Manikin Test [13, 34] (see Figure 2.3) is, as stated by its authors, "a non-verbal pictorial assessment technique that directly measures the pleasure, arousal and dominance associated with a person's affective reaction to a wide variety of stimuli." This test has its foundations on the Semantic Differential Scale, by Mehrabian and Russell [74]: a set of 18 bipolar adjective pairs devised in order to assess the three-dimensional structure of a variety of situations or concepts.

The test is age and psychopathology dependent, but was proven quite precise when judging the general emotional state of a subject sample after presenting a series of emotive stimuli to them [13].

The three 9-point scales of the SAM test were conceived as follows:

- **Pleasure:** Measures the level of happiness and pleasurable emotions present in the subject. A low score implies unhappiness, lack of satisfaction, melancholy or boredom.

- **Arousal:** Gauges the level of stimulation or excitement of the subject. A low score does not necessarily have negative implications, as it can mean calmness or relaxation, but also dullness or sleepiness.

- **Dominance:** Takes into account how influential the stimulus is over the subject. A low score means the person feels small or irrelevant in comparison with the stimulus, while a high score means the subject is in control, or in a situation of dominance.

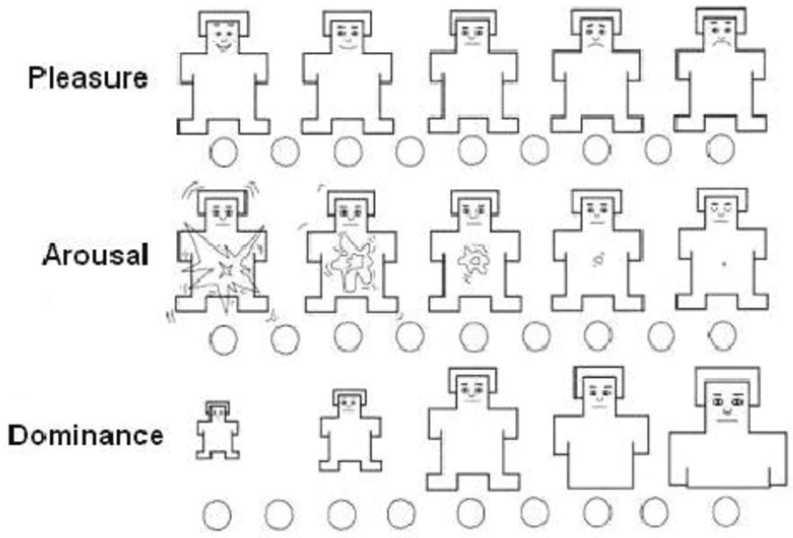


FIGURE 2.3. Original representation of the three scales in the Self-Assessment Manikin Test. Figure reproduced from [113].

This tool was utilised at a critical point in this research (see Chapter 7), as is detailed in Appendix D. Its capacities to assess all kinds of situations from an emotional perspective were useful when trying to understand the emotional state of players while listening to several audio tracks. The SAM test is suitable for audio tests in Spanish (which was the language spoken by all participants in this research process), and has been previously validated by Redondo, Fraga, Padrón and Piñeiro [94].

2.4 3D sound spatialisation in virtual worlds

3D audio is the use of a series of binaural recording techniques to capture, process and reproduce spatialised sound. 3D audio in video games is strictly related to presence, as it tries to reproduce an immersive sound experience with as much fidelity as possible [41], so that players can feel they are inside a given virtual environment.

From an academic perspective, spatial audio for video games has not been a particularly active or fertile field in the last few decades; the reason for this is that there were big conceptual advancements during the 1980s [82] which helped develop the technology we have today, and influenced the technical characteristics of stereophonic sound, surround sound and HRTF-based 3D audio, among others. Nevertheless, the popularisation of VR has brought a need to improve auditory experiences in video games [42], and this need has changed how we

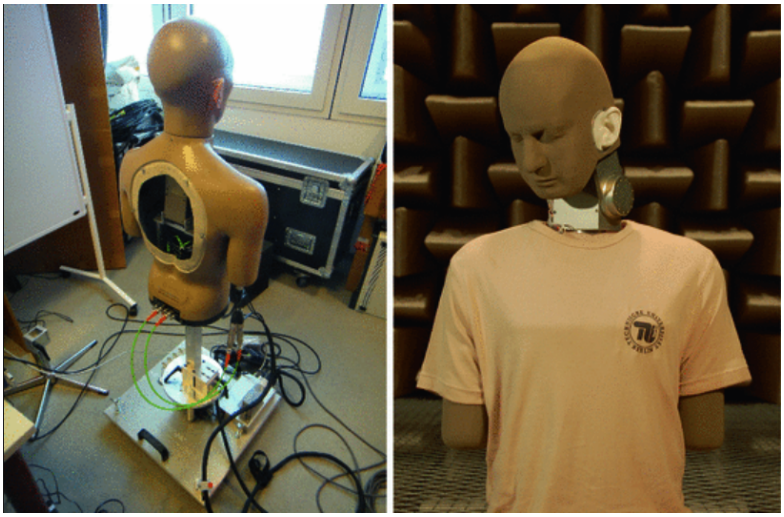


FIGURE 2.4. Dummy specifically designed for HRTF capture. Figure reproduced from [27].

conceive spatial audio for the first time in decades.

2.4.1 Head-Related Transfer Functions

The biggest recent advancement in the field of spatial audio is connected with the traditional concept of head-related transfer functions (HRTFs) [82]. HRTF sound capture is based on recording sound as it would have been heard by a real person in a real-world environment. To achieve this, several microphones must be placed in front of both ears, so that they capture sound coming from all relevant directions. The subject used for capturing does not have to be a human: it can also be a dummy specifically designed for HRTF capture, like the one depicted in Figure 2.4.

As an example, the KEMAR HRTF database [33] utilises one of this synthetic torsos to recreate the hearing conditions of a human auditory system. Other famous databases include: LISTEN HRTF [117], CIPIC HRTF [2], FIU DSP Lab HRTF [36] and ARI HRTF [4].

Information captured during an HRTF recording process is later used to create impulse responses, usually attached to audio events using a convolution reverberation plugin (like the one included in Ableton Live[4], for example). The result is a processed sound which can reflects the physical properties of the environment in which the impulse itself was recorded, and contains perceptible information about the physical direction from which said impulse came.

[4] https://www.ableton.com/en/packs/convolution-reverb/

When it comes to realistic audio content creation for video games, the combination of HRTF-based recordings and reverberation effects is the most commonly accepted option, because it allows for accurate 3D sound placement in a multichannel environment, as long as the final user has a compatible technology to reproduce it — a simple pair of stereo headphones should be able to replicate the effect.

The process of recording audio using HRTFs is, however, expensive and inconvenient for most real-life situations. This is why nowadays, the simulation of HRTF capture for video games is becoming very common.

2.4.2 Low pass filters

A more simple (and widespread) method for creating the illusion of 3D audio in video games is the utilisation of low-pass filters (LPFs). These are audio filters of a linear and time-invariant nature, which emulate a very common characteristic of sound in the real world: the attenuation or dissipation of high frequencies in a way that depends on the substance the waves travel through and the distance from the listener at which they are launched. This is achieved through these type of filters by simply cutting off frequencies above a certain number of hertzs (Hz) to avoid high-pitched sounds to pass.

The more *dissipating* an environment is, the more it cuts high frequencies. The dissipation factor (also called *absorption*) is increased mainly by two variables: the molecular structure of the medium itself (influenced also by temperature and humidity) and the distance a wave travels through it [123]. Human auditory systems are accustomed to perceiving this effect in daily life, and we use it as a clue to know where an object emitting sound may be located. For example: if we hear a sound which has many of its high frequencies cut off while at home, our brain tells us it may be in a different room.

Though the problem of simulating realistic audio absorption is complex, this phenomenon is easy to mimic in video games by attenuating the correct range of frequencies of a sound, taking into account its position and the physical properties of the virtual materials around it. A common way to achieve this effect is through a Chebyshev (type 1) LPF which, according to Williams [119], has an attenuation of:

$$A_{dB} = 10log[1 + \epsilon^2 C_n^2(\omega)]$$

Where $C_n(\omega)$ is a Chebyshev polynomial of the nth order that oscillates between ± 1 for $\omega \leq 1$.

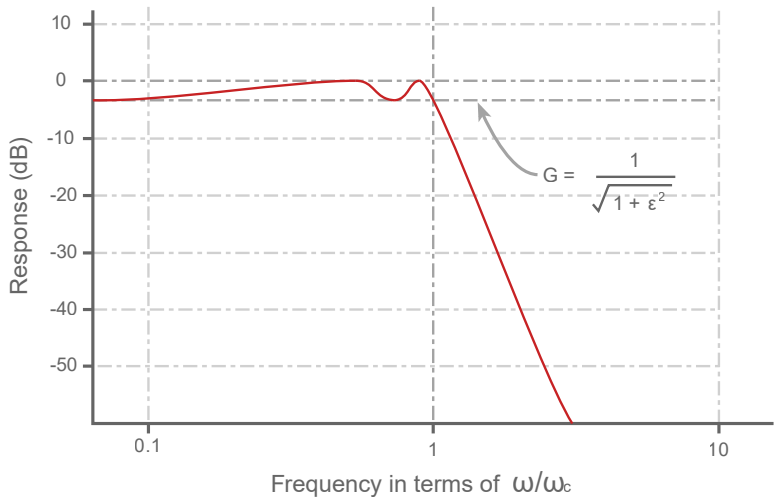


FIGURE 2.5. Graphical representation of a Chebyshev low-pass filter. Figure reproduced from [92].

$$\epsilon = \sqrt{10^{\frac{R_{dB}}{10}} - 1}$$

And R_{dB} is the ripple in decibels (dB).

Following research by Smith [107], type 1 Chebyshev filters grant a "faster roll-off by allowing ripple in the passband". This usually means the effect is quick, precise and clearer than other types, while maintaining a certain *smoothness*, as can be appreciated in Figure 2.5. Chevyshev LPFs are commonly used in the audio editing industry, and are one of the default options in popular software such as Adobe Audition[5].

2.5 Dynamic music and emotions

Virtual experiences, and particularly video games, are complex media in terms of sound design and music composition, due to the fact that there usually is no *linearity* —that is, a fixed succession of predictable events. Sound designers and musicians cannot predict what the exact outcome of a *gameplay* session will be, thus being unable to perfectly adapt a soundtrack to every game event.

[5] `https://www.adobe.com/products/audition.html`

2.5.1 Fundamentals of dynamic music

Because of this, dynamic music has found a place in the video game industry, and is still growing and evolving today as one of the most satisfying ways of creating soundtracks and audio events for games. Dynamic music, however, is a vague concept, and can refer to a variety of techniques.

Throughout the academic literature consulted while writing this thesis [15, 49, 109], the most common approach to dynamic music is to make soundtracks that are not strictly responsive (e. g.: as in a rhythmic game) but that adapt in some form to a series of events inside the game. As a consequence, ambient sound and music fit the general atmosphere of each possible situation, but do not adapt to particular actions instantly. Another common way of understanding dynamic music for interactive experiences is exemplified by the first two "Monkey Island"[6,7] games. In this case, dynamic music is utilised to ease the transition between scenes, and to avoid cuts in the soundtrack.

2.5.2 Problems of current dynamic music systems

The need to generate dynamic music to increase engagement in modern video games is a well known issue, and many companies in this sector have come up with solutions that usually entail big budgets and many resources destined to the music composition process.

For example, "Red Dead Redemption 2"[8], by Rockstar Games, utilises vertical re-orchestration, which consists of having a large pool of music stems, all of them recorded in A minor at 130 beats per minute in order to be easily combined.

"Tomb Raider Legend"[9] also developed the concept of micro-scoring: it adapted more than 180 minutes of music to different game characters, environments and events, and used a system which combined all the fragments in way that made sense from a musical standpoint.

The "Halo"[10] series, by 343 industries and Microsoft Studios, also features dynamic music: these games have a proprietary audio engine which combines vertical and horizontal composition techniques in real time, and randomises music so that it is slightly different every time someone plays the game.

All previous examples show that dynamic music is indeed valued in the video game industry; however, the methods described above are "handcrafted" and require a lot of time and resources. This makes them unfeasible for most companies. Besides, these techniques do not adapt completely to *gameplay* events, as they are based on making small variations

[6] https://www.gog.com/game/the_secret_of_monkey_island_special_edition
[7] https://www.gog.com/game/monkey_island_2_special_edition_lechucks_revenge
[8] https://www.rockstargames.com/reddeadredemption2/
[9] https://store.steampowered.com/app/7000/Tomb_Raider_Legend/
[10] https://www.halowaypoint.com

to an already prepared soundtrack, which is varied enough to provide the illusion of change and procedurality. Moreover, if players behave in a strange or unpredictable manner, music usually fails to adapt correctly due to a lack of time to come up with a believable transition.

In fact, the very concept of dynamic music is associated in academic literature with seamlessly changing between different pre-designed audio environments or scripted situations [109]. This means it does not adapt to player actions at all —only to their predicted behaviour.

Collins [15] also speaks of available technology as one of the main handicaps for truly procedural music in current video games. Procedural methods known at the time this thesis is being written produce low-quality results in terms of audio fidelity and musical quality, a fact that worries producers, as it could reduce player immersion, instead of increasing it. System performance also tends to lower considerably when using procedurally generated sound and music (and this is especially true if we aim for high-quality samples), which leaves less processing power for other important things like artificial intelligence and 3D graphics rendering. Moreover, long software development cycles in video games [93] and long console life cycles mean products are released with outdated technology in mind. Most popular mobile devices and portable consoles, such as the Nintendo Switch[11], also lack the power needed to generate high quality music in real time.

Some authors, like Jewell, Nixon and Prügel-Bennet [48] have developed systems for generating music by adding semantic information to genetic algorithms, which achieves better results in terms of musical quality. Even so, these techniques cannot be applied to real-time gaming with current hardware due to their elevated processing cost and the level of manual adjustment needed to achieve a sufficiently good soundtrack.

A similar problem affects other types of solutions, like the one proposed by Luhtala et. al. [71]. They add a wide variety of synth modules to an automatic composer, in a way that makes it possible to process raw MIDI data through Virtual Studio Technology (VST)[12] instruments, among other techniques. Realistic VST instruments, however, require high amounts of RAM memory, fast reading and writing speeds in the hard drive and a powerful CPU. High-end CPUs are not common in modern consoles (such as the PS4[13] and the XBOX One[14]), which makes this approach difficult to afford for many users. Even with a high-end PC, real time generation would be virtually impossible, as the rendering process takes minutes, instead of milliseconds.

Authors like Böttcher et. al. [10] also noted that there exists a generalised lack of attention to sound quality and music composition among video game players. This was also made evident through previous experiments [67]. Emotions frequently arise as an unconscious "side effect" of music, and Eladhari [25] also states that audio does not have an important role in a video

[11] https://www.nintendo.com/switch/
[12] http://ygrabit.steinberg.de/~ygrabit/public_html/index.html
[13] https://www.playstation.com/en-us/explore/ps4/
[14] https://www.xbox.com/en-GB/xbox-one?xr=shellnav

game unless it conveys important, gameplay-related information. As will be explained in the next subsection, however, players are indeed affected by the quality of a soundtrack (though indirectly), and the lack of appreciation of dynamic music does not mean there is no relation between its inclusion in a game and the level of presence achieved during the experience. Players explicitly asking for better graphics —and not for better music—, however, can influence how big companies release their products, and how many resources are destined to achieving the goal of a truly adaptive music system for games.

In fact, Livingstone and Brown [63] defend the need of a sound system which dynamically adjusts music in video games, taking into account player emotions while they play. The transmission of a personalised emotional tone for each player would give every experience a certain "cinematic" feel, and this may improve the perception of music in games most users have.

2.5.3 Basic emotions and dynamic music

Although the academic basis for understanding human auditory perception is ample, the interest in this field has been renewed recently. Authors like Asutay et. al. [3] have designed evaluation methods which consider the emotional meaning attached to a sound, along with signal characteristics, as a very important factor when identifying or recognising audio.

I. Ekman [23] indicates that designing audio for interactive experiences has to be in a strict relationship with the non-linear essence of modern media, and implies supervising and increasing emotional responses from users through real time adaptiveness. This author explains that there exists a "readiness to act" at the center of every emotion; thus, players are prepared for certain types of actions when feeling relevant emotions. Orchestrating the flow of these feelings should help the task of guiding or modifying player behaviour in an interactive virtual world. A good example of this would be a video game sequence in which the player has to flee from an enemy: if music urges users to do so, the action of fleeing itself would be understood as completely rational, and players would be prepared (and more inclined) to do what game designers thought optimal for this particular sequence.

During some of the research that led to writing this thesis [68–70], the need to have a simple categorisation for classifying emotions felt during a regular gaming session arose. P. Ekman's taxonomy of basic facial emotions [24] was proven useful for this task, as it only contains the 6 most basic human emotional reactions, and these can be easily identified in players during a gaming session [18]. The following list includes these 6 basic emotions:

- Anger: In close relationship with wrath or ire, produces displeasure and a certain inclination to belligerence.
- Disgust: Implies repugnance, aversion or distaste towards something or someone.

- Fear: It is described by P. Ekman as the urgency to avoid a painful, damaging or disagreeable situation. Humans can also fear an anxious reaction, even if the situation generating it is not potentially hazardous by itself.
- Joy: The feeling of being happy, pleased, fortunate or glad. It can have two dimensions: energetic or relaxing.
- Sadness: It means sorrow, grief, pain and lack of joy. Sadness is frequently associated with negative or bad things, and can be a consequence of fear.
- Surprise: A feeling that is sudden or unexpected. It has to do with learning by chance and discovery, and it can have a positive or negative tone.

These basic emotions can be easily induced through sound [30], which makes them ideal for the purpose of classifying different emotional reactions during *gameplay*.

Besides, authors like Yanagisawa and Murakami [122] state that it is possible to quantify the emotional quality of a sound (named "*kansei* quality" in said research) with parameters like "strong", "dull", "hard" or "silent". This knowledge would potentially allow a composer to increase or decrease the intensity of the level of joy a certain track is producing by tweaking said parameters.

Jørgensen [49] proposes a layered audio system in order to give enough emotional information to players. On one hand, a layer has to convey the general feeling of the virtual environment, and acts as ambient or atmospheric music. Other layers should offer players particular pieces of information about the game they are playing (e. g.: a stinger can play when there is a scare). In short: a layer sets the general mood, and additional layers include more detailed emotional information which adapts to certain *gameplay* events.

Most current dynamic systems take emotions into account. However, they tend to rely on scripted or predesigned sequences of actions, and the changes in tone and emotion are merely environmental. All the sound we perceive while playing one of these titles is what their creators thought optimal for a variety of situations or combinations of actions, but player emotions are not being monitored in real time. The development of emotional player analytics could help develop audio engines that behave in a really adaptive way in terms of emotion generation, and that overcome the difficulties present in current dynamic music systems.

2.6 Emotion extraction from textual inputs

In this context, "sentiment analysis" can be defined as an automated recognition of emotions present in a text. Resources in this field of Natural Language Processing are often used to find predominant feelings in big text corpora, or opinions given on social media by thousands

of people [88]. However, sentiment analysis has a much more uncommon facet: it can also be applied to situations during which a real time analysis of small portions of text is needed.

Most current techniques, however, do not offer the possibility to analyse complex emotions in a text. The most common libraries for sentiment analysis, like Cloud Natural Language by Google[15] or Natural Language Toolkit (NLTK)[16], tend to give simple and brief insights about the nature of the input. Said insights are normally related to two wide variables: polarity [89, 114, 116] and subjectivity [62, 90].

- Polarity is a variable that measures the general tone of a text. The output can be positive, negative or neutral in emotional terms. A positive outcome is associated with happiness, joy and agreeable feelings, whereas a negative outcome denotes sadness, anger, horror, disgust or a negative attitude in general. A neutral outcome is usually achieved when the artificial intelligence cannot determine the emotional nature of the text, or when it contains contradictory emotions.

- Subjectivity indicates if a text is objective or subjective, that is: if it gives personal opinions or contains purely objective statements.

This makes these text analysers optimal when, for example, detecting positive or negative opinions in online product reviews, but they fall short when trying to understand the actual emotion associated with the message (e. g.: if users are "scared", "angry" or just "sad" when receiving a negative input from them).

However, during this research, the different approach by Uros Krčadinac [53, 54] to sentiment analysis has proven very useful, as it goes beyond polarity and subjectivity. This author works from a lexicon-based perspective and takes into account Paul Ekman's basic emotions [24]. This means the output Krčadinac gets is richer from an emotional perspective, and can be useful in different case scenarios.

2.7 Player navigation and dynamic music

Making music useful for improving presence for video game players is a complex task, but an important one. Traditionally, music has been used to reinforce the general atmosphere of audiovisual products, but rarely has had a strong and clear functional value. Nevertheless, in current first-person and VR video games, music can have a determinant role as a tool for game designers, and this section explores one of the most immediate needs in this field: player orientation or guidance through auditory stimuli.

[15] `https://cloud.google.com/natural-language/docs/sentiment-tutorial`
[16] `http://www.nltk.org/howto/sentiment.html`

2.7.1 Player guidance in virtual worlds

A very common video game design challenge is related to the scope of possible player decisions. For game designers, it is a daunting task to guide players in virtual worlds that are bigger every day [108], and that contain an enormous variety of places to visit and things to do. Previous research [79] has found solutions for determining player goals by reading low-level information about *gameplay*; however, imposing behavioural tendencies through game design usually means introducing invasive techniques which reduce presence and break immersion. For example: telling players where they have to go by including a flashy marker on the graphical user interface (GUI) is effective, but immersion is maintained better if they walk to said location on their own. Besides, as is explained in Chapter 4, VR games should not make an excessive use of GUIs, as they reduce presence and can make players uncomfortable [55].

Thus, the problem of guiding a player through a big and complex virtual world is often solved by adding extradiegetic information to the GUI, affecting both presence [5] and immersion [46]. Certain video game genres can complicate matters even more: it is easier to justify the overabundance of GUI elements in a science fiction first-person shooter than in a medieval action game where the player should not have access to a navigation system.

In an attempt to standardise guiding techniques in virtual worlds, Milam and El Nasr [78] have established a taxonomy of behavioural design strategies in 3D video games, but audio-based techniques are absent from it, and academic work regarding this subject is scarce. This situation, far from being problematic, opens an unexplored field of possibilities for game designers, who often oversaturate GUIs due to a lack of alternatives (see Figure 2.6).

Non-graphic guiding techniques, in spite of not being common in the audio field, are not difficult to find in other domains, like narrative. Early research by authors like T. A. Galyean [32] relies on paths established through a narrative, which users must follow if they want to keep up with the flow of events. Recent video games like "The Stanley Parable"[17], "Dear Esther"[18] or "Gone Home"[19] all rely on narrative elements, like the voice of a narrator, to guide players to a particular milestone or goal. These techniques, however, only work in small, linear and highly controlled environments, and they need some form of textual narration to be happening (e. g.: a voice over, fragments of text scattered throughout the game's world, etc.). Open-world games that do not rely on text could benefit from the approach proposed in this thesis, due to its more subtle nature.

[17] `https://stanleyparable.com/`
[18] `http://www.thechineseroom.co.uk/games/dear-esther`
[19] `https://gonehome.game/`

FIGURE 2.6. Graphical user interface (GUI) in World of Warcraft. Figure reproduced from [1].

2.7.2 Auditory preference and meaningful variations

Authors like Eisenberg and Forde [22] show that it is possible to establish a series of simple predictors, like creativity, complexity or technical goodness, in order to explain the variations in preference during a human evaluation of music. Musical genres are the usual unit of measure when it comes to auditory preference or taste in humans [35], but modifying simple auditory features allows for a more flexible approach that works well with an adaptive music system like LitSens. Complexity, pitch and rhythm were the three modifiable attributes chosen during this research. This decision is consistent with previous uses of musical complexity (in and out-of-key notes, harmonic versus dissonant layering) [28], pitch (high and low tone, tuning) [73], and rhythm (slow or fast tempo) [17] to produce significant changes when playing an audio track.

Thus, an increase in complexity can be achieved simply by introducing layers of sound (formed by out of tune intervals and dissonant chords) that disrupt the harmony of a base track; pitch and rhythm, on the other hand, can be modified in real time inside the game engine.

As for what makes a sound "stand out" over others, a very popular opinion, based on classic works by Fletcher and Munson (the Fletcher-Munson curves) [31] is that sounds with a higher pitch —between 2000 and 5000 Hz— will usually dominate a mix in terms of perceived loudness. However, subsequent academic research clarified this issue: listener's perception of

several auditory attributes, including tone dominance, can be influenced by many different factors. Regarding pitch perception, in certain conditions outlined by Ritsma [95], lower frequencies can be dominant. And dominance is, in the context of this thesis, a determinant factor when identifying and following spatialised sounds in 3D virtual environments.

2.8 Artificial intelligence and music generation

One of the most popular ways to automatically or procedurally generate sound and music is to use artificial intelligence (AI) or any sort of automatic generation system. At this point, it is important to establish a clear differentiation between several types of computer-generated music, ordered by their level of *procedurality*:

- **Procedural music produced live. Output: audio file or stream.** This is the ideal scenario, where a musical track is composed and rendered in a convincing manner with a total delay of less than 42.32 milliseconds [126]. Currently, there are no systems with these characteristics in the reviewed literature.

- **Procedural music produced offline. Output: audio file or stream.** On the other hand, offline generation of fully procedural music is a more common practice. Companies like AIVA[20] can offer convincing, fully rendered audio files composed from scratch by their AI. Their approach is fragmented, however, and relies on multiple systems to generate a full piece of music: one system produces melodies, another one designs an accompaniment, and so on. This situation moves these kind of solutions away from real-time generation, as several, complex and time-consuming iterations are needed to achieve good results. Jukedeck[21] has a very different approach but reaches similar results: it generates symbolic scores that are translated into auditory information through synthesis. Jukedeck does not rely on MIDI to produce the final render like AIVA; instead, this technology uses machine learning during the whole synthesis process.

- **Semi-procedural music produced live. Output: MIDI file, audio file or stream.** This category can be subdivided in two groups: systems which produce MIDI files live (like the ones proposed by Devin Roth [98]) and systems which recombine existing audio fragments to create new songs. The adaptive music system proposed in this thesis would fall into this last category. These systems tend to work well and produce convincing results, but they need human intervention. MIDI files usually have to be tweaked and

[20] `https://www.aiva.ai`
[21] `https://www.jukedeck.com/`

rendered using VST instruments, which makes the whole process non-immediate and unsuitable for a real-time scenario; fragment-based compositions can be rendered in real time with ease, but a human composer must have previously made those fragments available to the machine, with a particular style in mind.

Additionally, one of the main problems that arise from live generation of MIDI information is that, if good results are to be achieved, a human must have manually prepared the software so that its activity is limited to a certain harmony or type of melodic progression, as can be seen in Figure 2.7.

```
var piano:[[Int?]] =
    [
    [nil,    nil,   nil,    nil,    nil,    nil ],
    [nil,    nil,   nil,    nil,    5,      nil ],
    [nil,    nil,   nil,    nil,    6,      -1  ],
    [nil,    4,     nil,    nil,    1,      nil ],
    ...
    [nil,    4,     -1,     nil,    3,      -1  ],
    [nil,    nil,   nil,    nil,    nil,    nil ],
    [nil,    3,     nil,    3,      -2,     1   ],
    [1,     1,     1,      nil,    3,     -1   ],
    [1,     1,     1,      3,      -2,     1   ],
    ]
```

FIGURE 2.7. Delimitation of MIDI notes for piano so that every note is in harmony in spite of randomisation. Figure reproduced from [99].

– **Semi-procedural music produced offline. Output: MIDI file, audio file or stream.** Most current procedural music systems work this way. The AI is usually trained with pieces from a particular genre or style, and then produces a MIDI file that can be later processed through VSTs. One of the best-sounding examples in this category is Deepbach [37], which uses machine learning and was trained on chorale harmonisations by the Baroque composer J. S. Bach; this system produces very convincing MIDI sequences resembling said author's style. A similar approach is taken by Nayebi and Vitelli [83], who use recurrent neural networks to achieve convincing note sequences that share the features of the training database. A different solution is given by Elias Software[22], which acts as an assistant or "arranger" for adaptive music by providing relevant fragment combinations to a human composer.

– **Non-procedural computer-generated music. Output: audio file or stream.** Non-procedural computer music is usually generated using software synthesisers or sampled instruments. A human composer must be present at all times.

[22] https://eliassoftware.com/

Markov chains [86, 87] are one of the most popular mathematical tools for generating procedural sequences of notes, and consequently, procedural music. A frequent use for this tool is tracking the probabilities of certain state transitions of a closed system, thus being able to predict behaviours. Music is made of complex, yet formalisable rules, and depends on frequent practices; if fed with a wide variety of note sequences that follow a particular style or genre, Markov chains would hypothetically be able to "autocomplete" series of notes with the most probable options.

Several authors [16, 39, 58] have achieved relative success when testing or implementing systems which use Markov chains as their main tool for note prediction. However, the problem with these approaches, as was mentioned in section 2.5, is that audio rendering itself can be very costly in terms of system resources, and those resources are already scarce if a video game is running at the same time as a procedural music system. The only scenario where Markov generation would be efficient enough to be applied to a video game is when using a standard like MIDI, being rendered through an interpreter like `IDirectMusicSynth` on Windows or `jMusic` on Linux.

Considering all this, a semi-procedural system would be the ideal option for generating high-quality audio while maintaining enough adaptability. This was the approach taken in this context, as will be explained further in later chapters.

CHAPTER 3

MOVEMENT AND PRESENCE IN VIRTUAL REALITY

"Distance in a straight line has no mystery. The mystery is in the sphere."

Thomas Mann

One of the first endeavours of this research was to grasp the true meaning of presence in virtual reality experiences, and its relationship with detrimental factors such as simulator sickness (SS). Even though the focus of this research was reaching higher presence levels through auditory techniques, the existence of a limitation such as SS when allowing for free locomotive movement in VR could have halted the progress of said research, and influenced the advancements described in subsequent chapters.

The problem of SS in virtual environments is complex, and it is still being addressed from a wide variety of perspectives in recent academic studies [29, 81, 91, 111]. After analysing both academic publications and industrial trends in video game development for VR, an experimental approach, described in [66], was taken in an attempt to clarify how presence decreased with artificial movement, and if it was possible to include several of the most popular locomotion techniques in a video game without making players sick or unengaged.

3.1 Experimenting with movement in VR

The core of said experiment consisted of two 3D VR video games which included the same goals but utilised different locomotion techniques. These two virtual environments were designed in accordance with Sherman and Craig's criteria about the essential elements of VR [104]. These authors talk about four key elements which must be contained in any VR experience: virtual world, immersion, sensory feedback and interactivity. Those elements should also be contained in a virtual reality system, that is: a combination of software and hardware which has inputs (monitoring users and the world around them) and outputs (visual, aural, haptic and vestibular displays). The best commercial hardware available at the time of this research was the Oculus Rift Development Kit 2, and the virtual worlds were designed specifically for it, while containing all mentioned elements.

3.1.1 Experiment design

Throughout this simulation, the act of *moving* or *walking* had many similarities to what Usoh, Arthur, Witton et. al. call *flying* [115]: locomotion along head direction, without real-world walking movements such as head bobbing, small accelerations and decelerations and changes in altitude, among others.

Said movement was controlled by the left analog joystick of a regular Xbox 360 controller. Forward movement followed the direction the player's camera was facing (its forward vector), and it was also possible to move sideways (often called *strafe*) and backwards (by inverting the forward vector). A combination of two of these four axis was also possible: the player could move, for example, 30 degrees to their left by combining the forward and left axis, while still facing forward.

Camera speed was similar to that of a human being while jogging (approximately 2 m/s). Initial acceleration was very quick (around 15 m/s^2), so that constant motion was achieved almost instantly after pressing the axis in a certain direction. This kind of setup is very common in most first-person video games.

As was explained before, two slightly different video game prototypes were built with the intention of measuring the effect of walking (or, in this case, *flying*) on presence and SS. Both environments revolved around a simple puzzle which had to be completed so as to trigger the end of the experience. However, there were differences in the distribution of in-game elements, and actual player displacement was only needed in one of them.

Both prototypes were made in Unreal Engine 4[1] from scratch, though several free assets from the video game "Infinity Blade" were utilised in the process[2] (these were made available

[1] `https://www.unrealengine.com/en-US/what-is-unreal-engine-4`

[2] `https://www.unrealengine.com/marketplace/assets?lang=&q=infinity%20blade`

for use with this engine by their developer: Epic Games).

The above-mentioned taxonomy of VR elements [104] was applied in the following manner to the design of each prototype, as seen in [66]:

- **Inputs:**

  — **World input and persistence:** The virtual world was responsive and behaved in a realistic manner. The player could collide with walls and objects, and a physics engine was implemented, so that smaller objects could be moved thanks to the force of gravity and the impulses coming from the player.

  — **User input and monitoring:** Input from the player was collected by using two devices. On one hand, the HMD built-in positional tracker was used to track the head; on the other, a Xbox 360 gamepad was used to track player movement.

- **Outputs:**

  — **Visual outputs:** An Oculus Rift DK2 was used as the visual output during both experiences. Its screen has a resolution of 960 x 1080 pixels per eye, and a maximum refresh rate of 75 Hz, which was stable in both experiences.

  — **Aural outputs:** Every user wore a pair of headphones (JVC HA-S400-B) during the experiment. This was the only source of aural output. The prototypes included both ambient sound and responsive sound (when the player stepped over a metallic object, for example).

  — **World representation:** The world had a low abstraction level, and the user could easily face all visual elements required for the completion of the puzzle. Both experiences took place in the same virtual world, the only difference being the position of key elements.

While wearing an HMD, one of the participants in the experiment was asked to identify all actions and elements present in each prototype. This decision was made to ensure that the noticeable core elements were the same in both experiences. This particular subject was 25 years old, male, and had a background in Computer Science. Besides, he had no knowledge of the nature of the experiment beforehand. The actions and elements he was able to identify were the following:

- **Game actions:** move, look at, interact and listen.

- **Game elements:** walls, floor, rocks, ceramic objects, metal objects, windows, ceiling, pedestals, orbs of three different colours, statues, fire, fog effects, game sounds and ambient sounds.

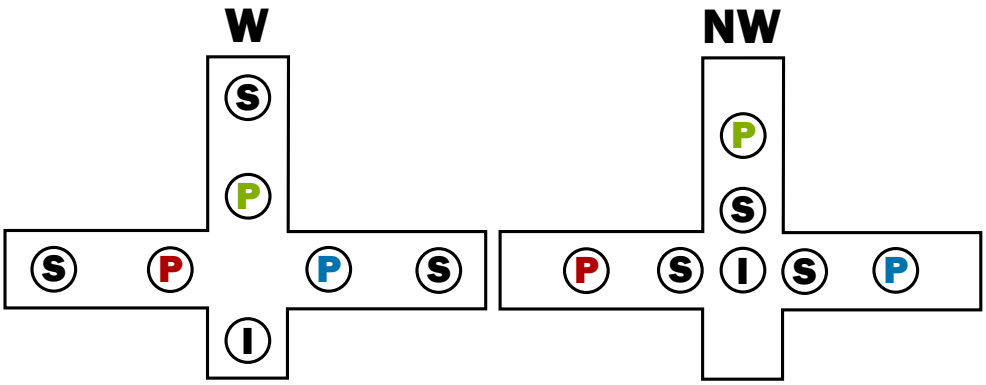


FIGURE 3.1. Schematic representation of the layout in each prototype. S represents a statue, P, a pillar and I, the spawn point.

FIGURE 3.2. First-person perspective view of a section in prototype W.

The first variation of the prototype (W) consisted of a 3D environment, navigable using the aforementioned HMD and an Xbox 360 controller. Figures 3.1 and 3.2 show the general layout of the environment. In it, a puzzle had to be solved in order to complete the experience, and this could be done by following a series of steps [66]:

The player had first to walk towards one of the three statues (S), located behind a pillar (P) on top of which there was a lit orb. This orb could be red, green or blue, depending on its position. Once in front of one of the statues, the player had to extinguish a fire that the statue held in her hands by pressing the 'A' button in the controller (an in-game tooltip indicated which button to press). The correspondent orb would also turn off as a result, leaving only the other two lit. The player had to repeat this process until no orbs were lit. Then, a gate would open in the place where the player first spawned (I). If the player stepped through it, the experience ended, and the puzzle was considered resolved.

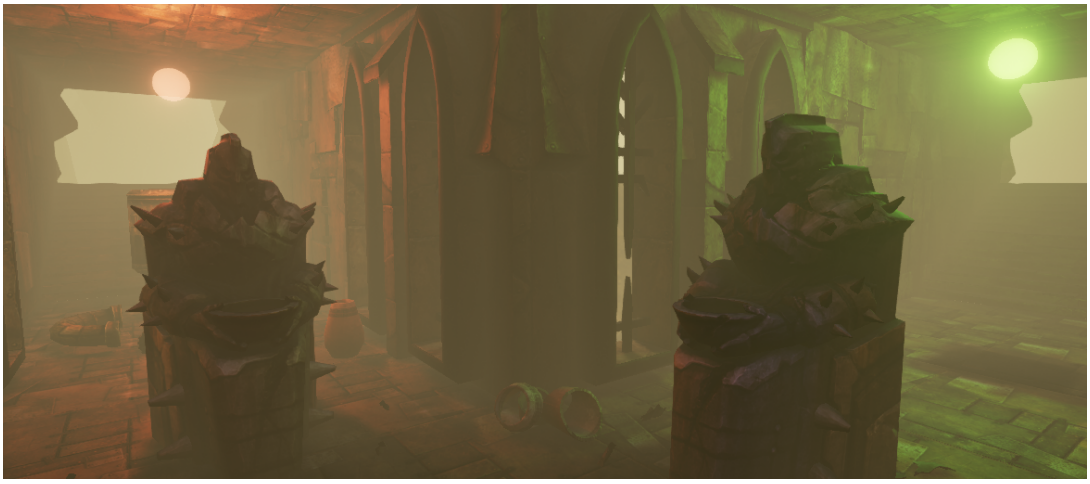


FIGURE 3.3. First-person perspective view of a section in prototype *NW*.

The second variation of the prototype (*NW*) had a very similar design, and contained the same kind of puzzle. However, every interactable game element was distributed in a way that made them accessible from the player's spawn location. In this case, the player was not able to walk (or *fly*); instead, they could only displace their head. This movement was not executed through a controller, but only by tracking each subject's head movements with the HMD, and the in-game camera merely reproduced said movements. The solution to the puzzle was adapted to this new situation: players did not have to walk anymore towards each statue, but they still had to turn off the three fires to trigger the end of the game.

3.1.2 Measures

The following measures were taken during the experiment:

- Presence questionnaires: Two Python applications were made to automate the process of passing the Temple Presence Inventory (TPI) [64] and the Slater-Usoh-Steed Questionnaire (SUS) [106] to every survey respondent. Results were also normalised (over 1) so as to be able to compare them later on. These surveys were passed as soon as the subject finished playing both prototypes and had answered the SSQ.

- Simulator Sickness Questionnaire (SSQ): Another Python application was designed to pass the SSQ [50], which was used to measure the level of simulator sickness each user felt after the experiment.

- Completion of the in-game puzzle: Subjects were told they could interrupt the experience if they felt sick or uncomfortable. A variable stored information about which users had

finished the experience, and how many did leave it unfinished for any reason.

All the original questionnaires utilised during this experiment can be found in Appendix B.

While passing the TPI, only the "engagement", "spatial presence" and "perceptual realism" subscales were utilised. "Social presence", "social richness" and "social realism" were not applicable to this experiment, as there were no social interactions of any kind in it.

Participants did not know any details about the experiment in advance, and all of them were simply told to play the game, to interrupt the experience whenever they wanted, and then to answer some questions about it. The SSQ was passed first, so as prevent the effects of SS from fading away too soon. The SUS and TPI questionnaires came after it, in that order.

3.1.3 Demography

Our sample was composed by 12 subjects that tried both experiences, in order to be able to compare their reactions to the differences in them. 6 of them took the experiments in one order, while another 6 took them inversely. Moreover, none of the subjects who took part in this experience knew about the existence of a second one beforehand.

All participants had to sign up by answering a brief questionnaire which contained the following categories: name, age, gender, education, VR device familiarity, VR devices present at home and video game familiarity.

Most subjects were men (11, 91.67 %), whereas only 1 person was a woman (8.33 %). Their ages ranged from 18 to 26 years old. 8 of them also shared a background in Computer Science (66.67 %), while 3 (25 %) were studying a degree related to Video Game Development and 1 (8.33 %) had studied a degree on Graphic Design.

4 of the participants (33.3 %) did not have any familiarity with VR devices, and had never tested a HMD; the rest had used at least one VR system, once. However, 8 of them (66.67 %) did not own any VR device, and only 4 (33.33 %) had a Google Cardboard at home. Lastly, all of them answered they were familiar with video games, and declared themselves frequent players.

3.1.4 Results

In spite of the reduced sample, results point to the existence of a strong correlation between the action of walking and an increment in presence. The results were consistent for both SUS and TPI questionnaires; after variance analysis, the two sets of results presented a p-value

lower than 0.05, and they were easily comparable (as can be appreciated in Tables 3.1 and 3.2).

TABLE 3.1. Normalised SUS test results (One-way ANOVA).

Walking	**Yes**	**No**
N	12	12
SUM (X)	8.285714	7
Mean	0.690476	0.583333
Variance	0.010101	0.003556
***p*-value**	0.00413	

TABLE 3.2. Normalised TPI test results (One-way ANOVA).

Walking	**Yes**	**No**
N	12	12
SUM (X)	7.759259	6.092593
Mean	0.646605	0.507716
Variance	0.007386	0.004502
***p*-value**	< 0.0001	

To analyse these results, it is important to take into account that in-game movement did not imply real-world motion to any extent. This lack of correspondence between in-game and real-life movement does not seem to be detrimental for presence when the user *flies* in the virtual world.

Figures 3.4 and 3.5 show the numbers assigned to each subject in the abscissa axis, and normalised levels of presence achieved in the ordinate axis. Even though Figure 3.4 presents a more irregular shape than Figure 3.5, the results are consistent for both questionnaires. This difference in regularity may have been caused by the different number of questions in the two tests: this version of the TPI had 18, while the SUS had only 6.

TABLE 3.3. Normalised SSQ test results (One-way ANOVA).

Walking	**Yes**	**No**
N	12	12
SUM (X)	2.416667	1.0625
Mean	0.201389	0.088542
Variance	0.045718	0.006323
***p*-value**	0.120654	

On the other hand, the SSQ offered very similar results in both experiences, which was counter-intuitive. After a variance analysis, the two sets of results did not offer a significant correlation, and a p-value of approximately 0.12 was obtained (see Table 3.3. Thus, the act of

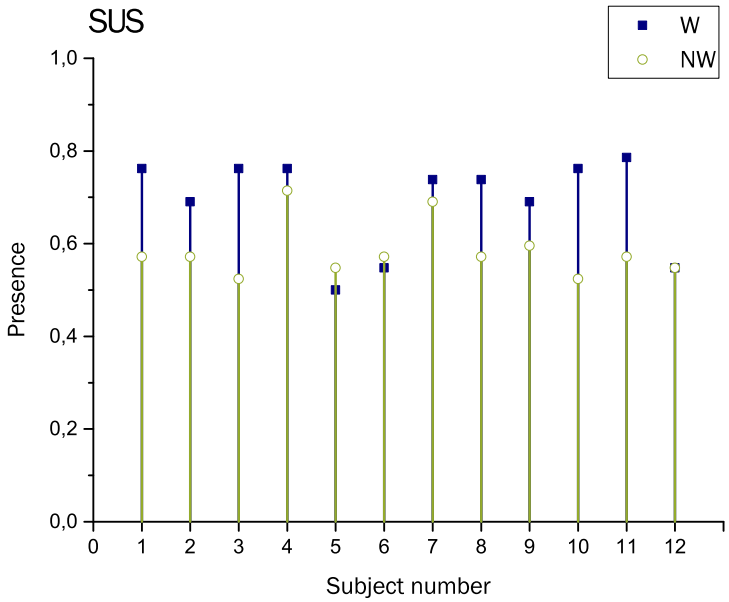


FIGURE 3.4. Normalised results of the SUS presence test for both walking (W) and not walking (NW) experiences.

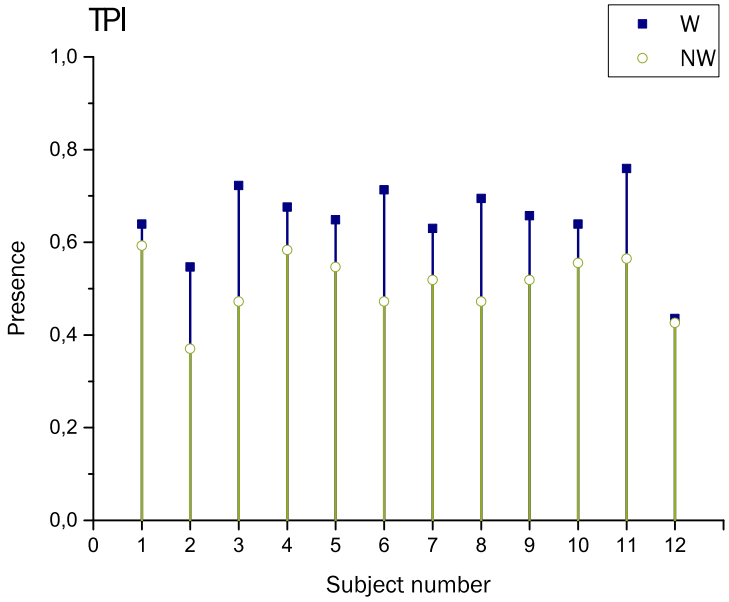


FIGURE 3.5. Normalised results of the TPI test for both walking (W) and not walking (NW) experiences.

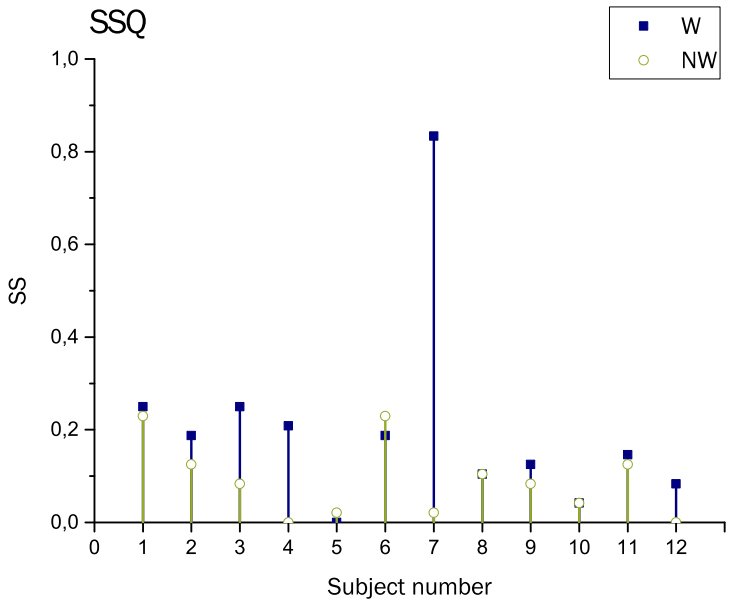


FIGURE 3.6. Normalised results of the Simulator Sickness Questionnaire for both walking (*W*) and not walking (*NW*) experiences.

walking did not seem to noticeably affect SS. Only subject number 7 (as can be seen in Figure 3.6) experienced an increment in SS while walking. However, this person had defined himself as particularly prone to nausea and motion sickness, even in first-person experiences outside of VR.

3.2 Conclusions

Due to the small amount of subjects who took part in the experiment described above, its results should be taken as an indicator of an inclination towards a certain type of behaviour —never as a strong deduction. However, the results of both SUS and TPI questionnaires were in accord, even though the experience was brief, and the SSQ points to a lack of SS in situations where players move. After this experiment took place, several companies have published high-budget VR video games with *flying* movement, and the results have been satisfying enough to dilute the taboo of VR locomotion. Some of these games are "The Elder Scrolls V: Skyrim VR"[3], "Apex Construct"[4] or "Red Matter"[5]. The developers of "Apex Construct" have even justified their decision to include locomotion in their game by popular demand. This method is not mandatory, as players can choose how they want to roam the world in this

[3] https://store.steampowered.com/app/611670/The_Elder_Scroll_V_Skyrim_VR/
[4] https://store.steampowered.com/app/694090/Apex_Construct/
[5] https://store.steampowered.com/app/966680/Red_Matter/

particular video game, but, as the developer Erik Odeldahl states [38]: "the numbers we're seeing are just way, way higher than we thought they would be. It turns out that, across all of the platforms we released on, 46 per cent played the game using free locomotion. It's a mind-bogglingly large percentage. Half the players of our game use free locomotion".

Besides, all of the subjects who took part in the experiment were experienced video gamers, which means they could be more used to virtual motion, and therefore more resistant to simulator sickness in short and focused experiences as those of our experiment.

In summary, it is possible to assert there is a significant difference between *walkable* and *non-walkable* first-person VR games in terms of presence and immersion. Letting the player move longer distances than what the tracking capacities of the HMD would allow, even when there are no real body movements associated with the act of walking of *flying*, generates an increment in the level of presence, granted by the sense of freedom it induces.

The results obtained also point to a relationship between SS and the design principles assumed when building the systems that allow players to move in VR. The action of moving itself does not seem to necessarily induce SS, but moving in a certain way does. The trends observed in the video game industry since this experiment was published reinforce this line of thought, since more and more *walkable* VR games are being published.

Once free locomotion was proven possible in VR without affecting presence or immersion in a determining manner, the next goal was to improve that movement by either guiding or reinforcing it through auditory interfaces. This whole process is detailed in the next chapter.

CHAPTER 4

THE INFLUENCE OF SOUND OVER GAMEPLAY

"Sound unbound by nature
becomes bounded by art."

Dejan Stojanović

To be able to influence how a player behaves inside a virtual environment is one of the classic challenges posed by the field of soundtrack composition for video games. The influence of VR in current game design, and the persistent need to create more immersive experiences, means graphical user interfaces (GUIs) cannot contain all the elements needed to understand the game flow in complex experiences, as this would go to the detriment of presence [55].

During initial research for this thesis, an experiment was designed to determine if sound could influence decision-making in video games, and how different psychological profiles behave when facing changes in sound. The preliminary hypothesis was that music could have a certain degree of influence over how players behave when facing binary decisions in a given virtual environment.

This experimental phase, along with the one described in the previous chapter, laid down the foundations for the advancements presented from Chapter 5 to 8. First, the mere ability to move freely in VR was put to the test, and this chapter describes how viable it is to modify movement patterns and player decisions in spatial terms only by including different types of music and audio.

4.1 Experiment design

A survey (see Appendix A) was passed to a total of 92 people of both sexes over the internet. The intention behind this survey was to recover information regarding age group, sociological and psychological profiles, video game playing routines, general opinions about sound and video games and capacity of being influenced by a certain type of music when making decisions in virtual environments.

As Appendix A shows, the survey included a central question: subjects were asked to choose between two pairs of doors in an imaginary environment, which had to be pictured mentally through a brief description.

There were two phases during which users had to make a decision (between door A or door B). During the first phase, subjects listened to a stereo soundtrack which contained several key differences between the two channels; the song that played during the second phase, however, tried to be more confusing and muddy.

4.1.1 First phase

The first musical theme tried to clearly communicate two different emotions, one for each channel: the left channel had to be safe, calm and light, whereas the right channel had to communicate feelings of discomfort, danger and sadness. The main features of these channels were the following:

- **Door A (left channel):**
  - — High-pitched tone, with an abundance of frequencies above 3 KHz.
  - — In harmony, lack of dissonant tones.
  - — Similar timbre and waveform found in all the instruments used.
  - — Constant intensity; the average wave amplitude is higher.
- **Door B (right channel):**
  - — Low-pitched tone, with an abundance of frequencies below 300 Hz.
  - — Chaotic style, with a strong presence of atonal sounds. Lack of harmony.
  - — Use of non-complementary waveforms and very different timbres, which generates a "noise" effect.
  - — Frequent changes in intensity; the average wave amplitude is lower, though sometimes it peaks higher than the left channel.

4.1.2 Second phase

The second soundtrack, on the other hand, had the intention of confusing subjects on an emotional level. Its main features were distributed in both channels, and can be summarised as follows:

- Tone polarisation: The left channel includes low frequencies (around 300 Hz), whereas the right channel contains highs and mids.
- The right channel contains noise that muddles the high tones and avoids harmony.
- Both channels play a clear rhythmic pattern.
- There are wide tonal changes in the main melodic line, which is distributed in both channels and jumps from one to the other.

No permutations were made to the order in which these phases took place for each user. The reason for this is that users could start identifying random musical attributes as important if facing the more confusing version of the mix first, thus affecting their opinions in the next phase.

4.1.3 Demographic data

The ages of the surveyed population were distributed as is shown in Figure 4.1.

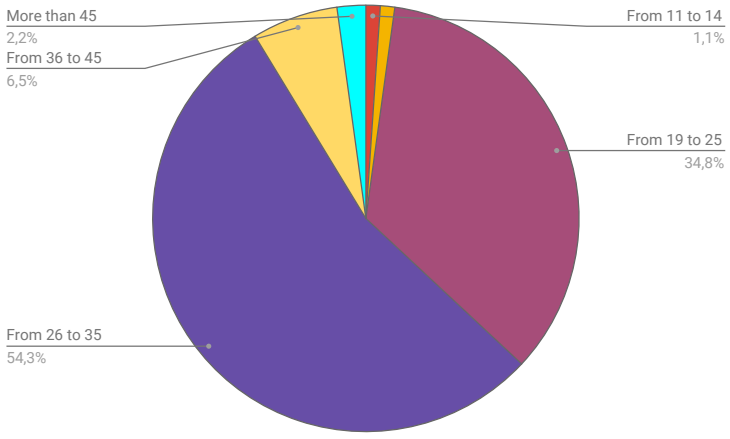


FIGURE 4.1. Age distribution in surveyed subjects.

Though the target subject group was general, and the survey was not oriented to any particular age range, participants were predominantly between 19 and 35 years old. This can

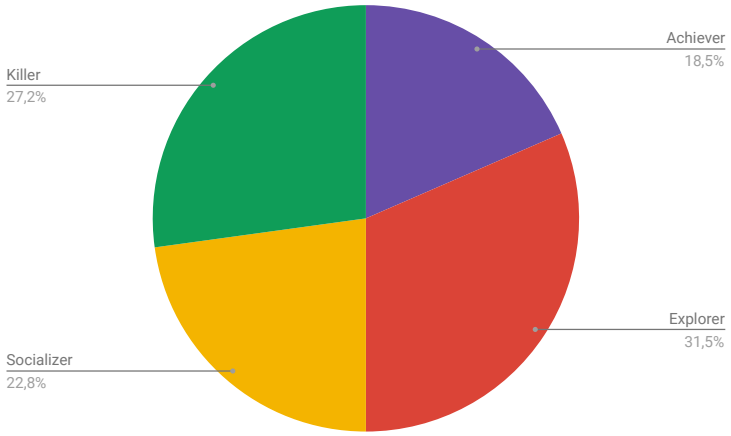


FIGURE 4.2. Psychological profile distribution in surveyed subjects.

be explained by the distribution of *gamers* in Spain, where this survey was taken. According to the GameTrack Digest (Quarter 3, 2018) [61], a 68 % of the Spanish population plays video games in the 15 to 24 age range; a 60 % keeps playing when they are 25 to 34 years old. However, only a 41 % of people aged 35 to 44 play video games, and the percentage drops to a 25 % for people who are 45 to 64 years old. Thus, the interest in taking a survey on interactive experiences would be reasonably higher for people between 15 and 34 years old.

Most survey respondents played video games frequently: a 41.3 % said to play on a daily basis, whereas only a 3.3 % never played games. A 39.1 % stated that their favourite genre is RPGs. This is consistent with recent studies [100] which explain that this genre is usually preferred by *hardcore* players, that is: people who play frequently and who have a profound knowledge of global video game production.

All polled subjects were anonymous, and did not know about the research beforehand. They were contacted through social networks.

The distribution of psychological profiles follows a regular, even pattern, as was expected from a general population. This can be appreciated in Figure 4.2. The percentage of subjects in each category is high enough to be able to analyse the sample with statistical rigour. There exists, nonetheless, a slight dominance of the "explorer" and "killer" profiles.

4.2 Results

The results obtained through the survey described above are consistent with the initial hypothesis, and point to a relationship between musical characteristics and player behaviour during an interactive experience. The answers to the question "Which door would you open

first?" changed depending on the soundtrack that was playing in each case. These results also varied for different player profiles.

4.2.1 First phase

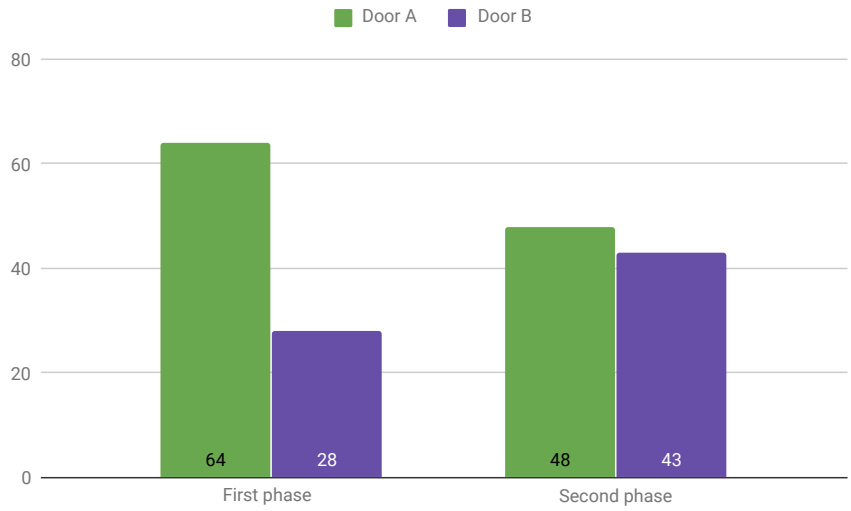


FIGURE 4.3. Door selection in both phases of the experiment.

As Figure 4.3 shows, most subjects were able to identify the position of both channels in space, and most of them chose the left door. Besides, 89.1 % of the participants situated the white door on the left, and a mere 10.9 % thought it was on their right side. Thus, the musical features of the left channel (high-pitch, harmony) are associated with the white colour, whereas chaotic and atonal melodies are associated with the red colour. These associations, according to Ou, Luo, Woodcock et. al. [85], are not exclusive of a particular culture, and happen in different countries around the world.

When asked what they think would be behind each door, most subjects answer with open and peaceful spaces for the white (left) door, and use keywords like "ocean", "wind", "peace", "forest", "blue sky", "angel", "Heaven", "sunny exterior", "an exit", etc. The red (left) door is associated with small and oppressive environments, described with words like "scare", "challenge", "rituals", "ambush", "enemy", "cataclism", "dark room", "hell" and "fire", among others.

If we take into account that most subjects knew which was the "safest" option, we can explain the fact that a 30.4 % opted for door B by the differences in psychological profiles present in the sample (as described above, in Figure 4.2). The results show that most people who chose this option were *explorers* according to the Bartle taxonomy. It seems logical that

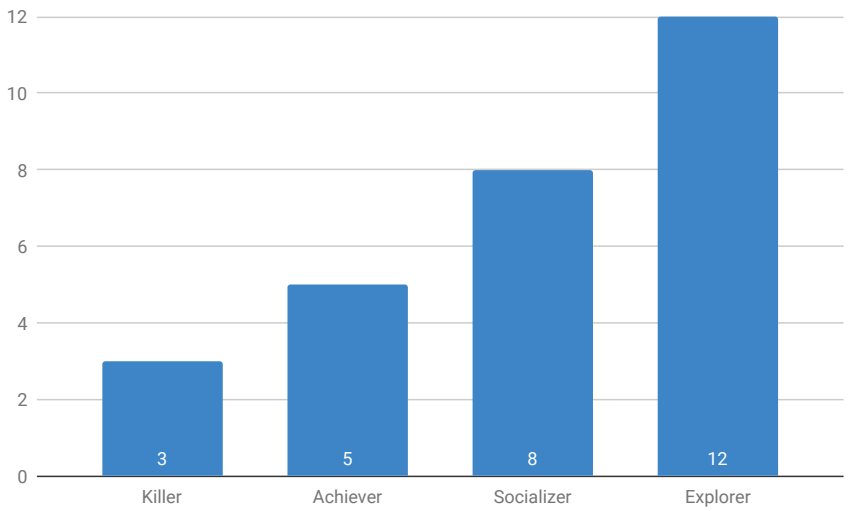


FIGURE 4.4. Subjects opening door B (red), organised by psychological profile.

this particular group chose as they did, considering their curiosity and their desire to know what happens in every corner of the virtual world. Figure 4.4 clearly shows this trend.

Socializers are also prone to experiencing new things and knowing the world around them better, which situates them in a second position in this case. *Achievers* and *killers*, however, do not find exploration interesting enough, mainly because their goal is to win and survive.

4.2.2 Second phase

During this second phase, results were less predictable, as subjects were not able to identify the position of each door.

Figure 4.5 shows a very distributed set of results. Without a soundtrack specifically designed to guide participants towards a particular option, we can assume a certain randomisation of choices.

When asked about why they changed opinions between phases, subjects answer mainly by describing musical features. They also point out that a change in the soundtrack also implies a change in genre (e. g.: from "adventure" to "horror"), thus producing a different decision. Several *explorers* mention that, during the second phase, the white door does not necessarily seem to contain a conclusive environment, and that it is difficult to know which option would guarantee that the player can explore everything before "getting out". They also say that they can sense a "hidden" or a "confusing" message, which makes it impossible to know with certainty what will happen after exiting through each of the doors. *Achievers* detect urgency in the soundtrack that plays during the second phase, which makes them choose the white

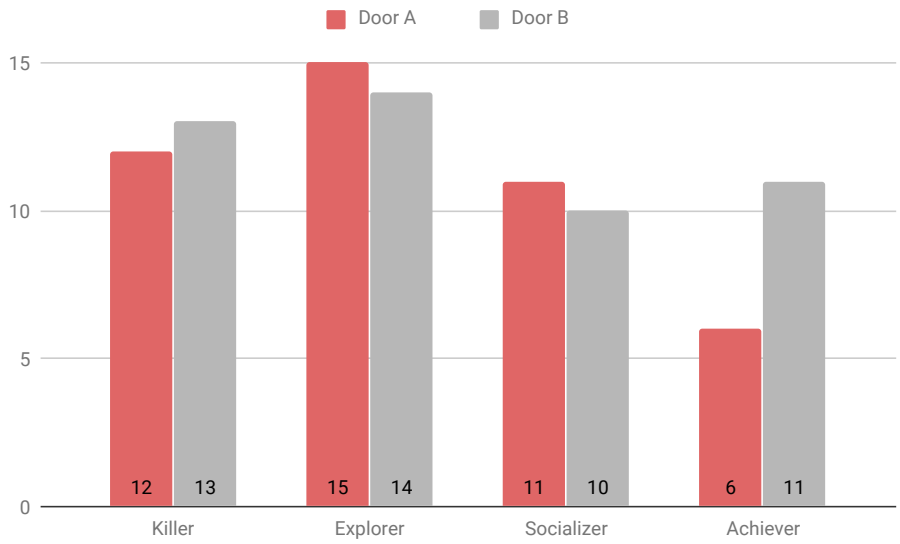


FIGURE 4.5. Door selection during the second phase of the experiment. The results are organised by psychological profile.

door (door A, on the left), because they remembered it was the "safest option" in the previous phase.

It is possible to conclude that the *correct* option was not clear at all for any group in the sample during the second phase, which means the soundtrack actually helped in the first phase, and was able to influence how almost every subject made decisions while listening to it. Not even the colour code can tip the scale towards a particular option in the second part of the experiment: about half the participants chose each given option.

4.2.3 Additional results

Survey respondents were also asked about the importance of sound and music in video games. On average, they valued music over sound, but these two elements were the last in a list that contains the following items, sorted by the importance given to them by participants:

1. *Gameplay*
2. Story
3. Graphical quality
4. Music/soundtrack
5. Sound design

Although sound and music can influence player's decisions, these elements remain unnoticed for the most part of a regular video game. Several participants stated, after taking the survey, that sound or music only stand out for them when there is something wrong with these elements or when they try to produce emotions that are inconsistent with the game's art style or *gameplay*.

When asked about the genre for which music and sound are more important, a majority among the respondents answered "horror". A 43 % selected this genre when talking about the soundtrack, whereas a 61 % thought scary games have to rely heavily on sound design.

4.3 Conclusions and design guidelines

Even though further experimentation was needed at this stage, it was already possible to extract a series of conclusions from the reviewed data.

On one hand, subjects were proven vulnerable to manipulation through the use of music, as long as certain rules are applied. A soundtrack can even determine which decisions are taken by a majority of players in the kind of environment described above, where the decisions are twofold and simple enough to be taken quickly.

Besides, participants seemed to be able to associate a colour (red or white) to a series of sound properties. The consistence of sound and graphics was also mentioned by several subjects as an important factor when making these kind of decisions in a video game.

Lastly, player's psychological profiles have to be taken into account when trying to produce a certain behaviour, as they seem to influence the final outcome in some cases.

Some general audio design guidelines can also be extracted from the observed behaviour in most subjects. These are mere hints which would have to be confirmed by further research, but may be of use when designing emotional audio for games:

1. 3D audio positioning using stereo techniques is useful when trying to communicate an object's location in a virtual environment.

2. High-pitched tones are interpreted as a positive sign in most cases, as long as they keep harmony. If they are inharmonious or atonal, they are perceived as a threat or as a reason for being scared or worried.

3. Harmony is associated with positive feelings, while inharmony is linked to negative emotions.

4. The combination of instruments with similar or complementary waveforms produces a feeling of peace and calmness in many users.

5. Sound in a virtual environment should be coherent with dominant colours so as not to produce perceptual dissonance.

6. The utilisation of a wide range of frequencies in both stereo channels produces confusion when trying to position an object in a 3D virtual world.

7. The introduction of noise reduces or negates the effects described in 2, 3 and 4, and induces feelings of unease.

8. A strong rhythm can produce a feeling of urgency, even if the harmony is strong.

9. Wide-range changes in tone produce discomfort and disorient subjects, unless executed very slowly. This is especially true if the intervals go beyond two octaves.

CHAPTER 5

IMPROVING SOUND SPATIALISATION IN 3D ENVIRONMENTS

"Silent the old town...
the scent of flowers
floating...
An evening bell."

Matsuo Bashō

After confirming the influence sound can have over how a potential player behaves or moves in a virtual environment, the next logical step was to analyse the relationship between several types of spatial audio and in-game performance. This research, described in [67], took place through two different experimentation phases, both of them aiming towards a better understanding of how spatial audio is perceived by subjects, and how it could be used as a game designer's tool. This would pave the way towards more complex approaches to spatialisation, such as the ones described in Chapter 7.

5.1 Hypothesis

The main hypothesis behind the two experiments that will be described in this chapter was that a low-latency spatialisation technique based on position-dependent LPF could allow for more accurate sound position identification when compared to a 3D sound system based on

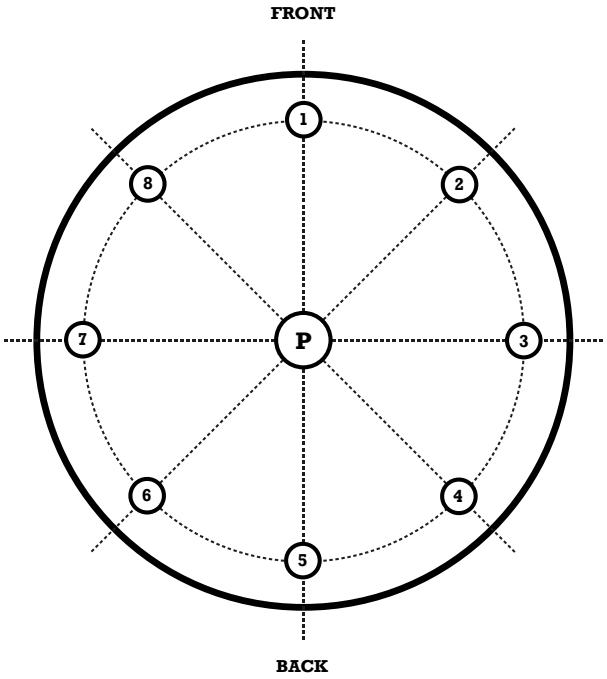


FIGURE 5.1. Position diagram for the first experiment. P represents the player, and numbers indicate all possible sources of sound.

panning, such as the default audio system present in most video game engines like Unreal Engine and Unity 3D.

The null hypothesis (H_0) was that performance of users when identifying sound position did not improve when using an LPF-based approach. The alternative hypothesis (H_1) was that performance improved only when using the LPF-based system.

5.2 Experiment 1: Online survey

The first experiment was conceived as a way of getting information about how well regular video game players perceive spatial sound, and how this could help them when positioning a certain object in a 3D virtual world. The experience consisted of an online survey, distributed through social networks, in which participants had to express their opinion on the position of a sound in a complex audio environment while using a pair of headphones. Said sound was a high pitched alarm tone which was played sequentially from 4 different directions in a virtual space, from a total of 8 possible positions (depicted in Figure 5.1. Users could not count on any visual references while taking the survey.

During this experiment, two different audio tracks were playing simultaneously: the

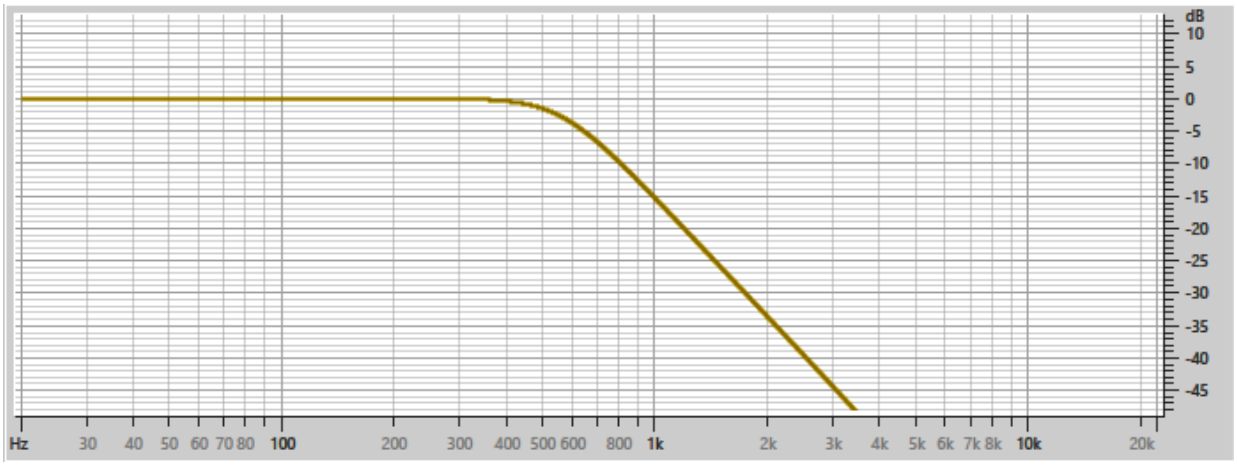


FIGURE 5.2. Attenuation graph achieved by applying a LPF in Adobe Audition.

aforementioned alarm sound was one of them, normalised at -0.45 dB; the other track contained a complex audio scene extracted from an action video game (a technical demo by Epic Games called "Showdown VR"[1]), and was normalised at -3 dB. The presence of both of these tracks forced users to detect and separate the alarm sound from a complex 3D sound environment which included sounds of guns firing, cries, explosions, metal impacts, etc., coming from a wide variety of directions.

Participants in this survey were separated in two groups of 29 subjects, and a different version of the test (1A and 1B) was passed to each of them. The first version (1A) contained the original sounds from "Showdown VR", along with four alarm sounds coming from diferent positions. Everything was played in Unreal Engine, while applying its default 3D spatialisation to each sound. The second version (1B) made use of the same background track, but the alarm sounds were processed in a different manner: a low-pass filter (LPF) was applied when they came from behind the player. The software Adobe Audition was used to process said audio tracks by applying Chebyshev LPF filters (see Chapter 2), achieving attenuation graphs similar to the ones in Figure 5.2.

A complete version of the test utilised for this experiment can be found in Appendix C.

5.2.1 Test structure

Excluding the differently processed alarm sounds, both surveys had the same structure, which was as follows [67]:

[1] `https://www.unrealengine.com/marketplace/showdown-demo`

– First, users were asked to listen to an isolated, non-spatialised sample of the alarm sound.

– Next, it was explained to them how to position sounds in a diagram like the one in Figure 5.1.

– Then, subjects were presented with a sound track which included four consecutive alarm sounds on top of the background noise described above. They were asked to identify their positions and mark them in the diagram. For example: first sound in position 4, second sound in position 1, etc.

– Lastly, a series of questions related to the demographic profile of each subject were asked. They included: age, sex, level of education, diagnosed audition (hearing) problems, perceived performance during the experiment, frequency with which the subject plays video games and opinion on the importance of video game audio.

All questions related to subject's opinions were posed by using a Likert scale [59].

5.3 Experiment 2: On-site test

The next experiment was an on-site test (detailed in Appendix C) during which participants had to complete a minimalistic 3D game made in Unity[2]. The utilisation of a consumer-quality pair of headphones (JVC HA-X570) was required. Besides, there were also two versions of this test (2A and 2B), which were taken by 13 subjects each. Identification tasks contained in this test were similar to the ones in the first experiment, but the experience took place in a playable 3D environment. Thus, this experiment was designed to recover higher quality data in a controlled environment that resembles a real video game.

The two variations of this application consisted of an empty room with a flat texture. Eight spheres (see Figure 5.3) appeared and floated around the player, following the positions depicted in Figure 5.1. All participants had to complete a brief interactive sequence during which they had to point and click the sphere they thought was emitting an alarm sound. The game was controlled with a mouse: its movements rotated the camera over a fixed position, and a left click activated one of the spheres. To help players point at things in 3D, a small crosshair was added to the centre of the screen. If the position selected through this method was correct, the sphere would stop playing its sound, start emitting a blue light, and the next sphere of the sequence would start playing the same sound from a different position. The game ended only when all of the spheres had been correctly selected and turned on.

[2] `https://unity3d.com/`

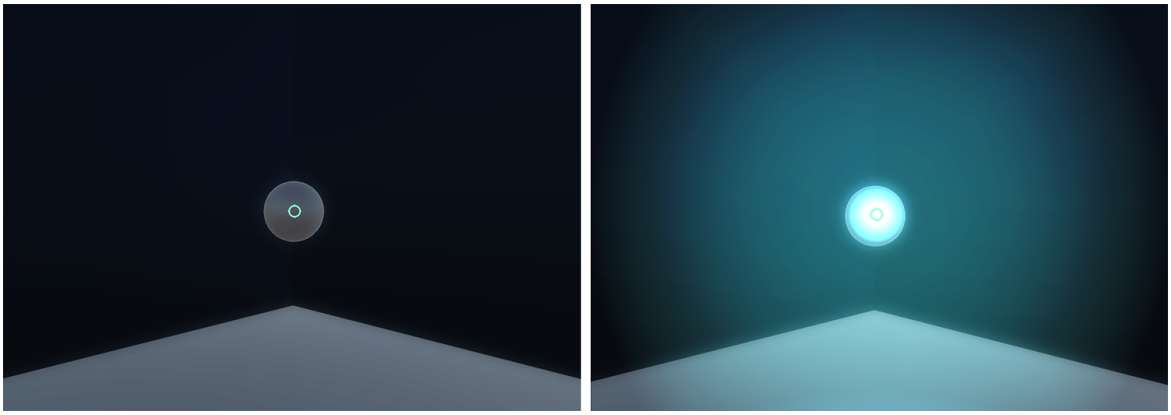


FIGURE 5.3. Two of the spheres present in experiments 2A and 2B. The left one has not yet been activated, whereas the one on the right has already been identified and clicked by a player.

This second experiment was conceived as a way to increase feedback for subjects, who knew how well they were doing as they advanced. Time was a very important variable in this case: it determined how well a participant did, while allowing everyone to finish the experience (there was no time limit).

All subjects received the same information before starting the experiment. The following instructions were read aloud to them, as is described in detail in [67]:

- You will play a game from a first-person perspective.
- You will be inside a small and dimly lit room.
- You will not be able to walk. You will, however, be able to look around using the mouse.
- A crosshair is shown at the centre of the screen. It indicates where you are looking at, and always follows the position of the mouse.
- Eight spheres will float around you. All will be at the same distance from you, and static. They will also be at the same distance from each other, forming a circumference around you.
- At the beginning of the game, one of the spheres will produce a looping alarm sound. Your task is to identify the sphere from which that sound is coming, point at it with the crosshair and click the left mouse button.
- If you identified the position correctly, the sphere will turn blue and the same alarm will start playing from a different sphere. If you failed to identify the position, the alarm will keep sounding until you do.

– The game will finish when all eight spheres are blue and you do not hear any more sounds.

– You must complete the task as soon as possible.

Data from every user was saved in a log, which contained raycast hits from the crosshair, total time to complete the task and time at which each sphere was clicked and lit. In this context, time served as a measure for performance quality, and times were compared between the two versions of the same experiment so as to know which one was completed more easily on average.

Additionally, everyone had to complete a small survey after finishing with the given task; it asked for the same demographic information as in experiment 1, but added several questions in which participants were asked to express their level of agreement with the actual location of the sounds.

5.3.1 Audio systems in experiment 2

There existed only one difference between applications 2A and 2B, and it was the audio system used by them. On one hand, 2A used the original 3D sound system from Unity (no plugins added); on the other, 2B utilised a custom sound system made specifically for this purpose. The latter system worked as follows. Each frame, all audio coming from visible spheres (that is, from spheres being rendered by the player's camera) was processed by using simple stereo panning and distance attenuation. These are the methods used in Unity by default to spatialise sounds. Sounds coming from objects not being rendered by the camera were applied a LPF. The parameters used for this filter were a cutoff frequency of 2456 Hz and a low pass resonance of 1. This effect is not necessarily consistent with how high frequency attenuation works in reality, but is clearly and easily identifiable.

The field of view (FOV) of the in-game camera tried to be as similar as possible to the frontal eye field (FEF) of human eyes: 114 degrees [44]. Everything outside this on-screen space was considered to be in the rear, with no distinction between "side" and "back".

5.4 Demography

Experiment 1 was offered to a sample of 58 people (41 men and 17 women), randomly distributed in groups of 29 persons for each version (1A and 1B) of the test. Average ages in groups 1A and 1B were 33.03 and 31.28, respectively, and ranged from 22 to 51 years old in group 1A, and from 20 to 43 in group 1B.

48 of these people had gone through college (17 had a degree, 29 a master's degree and 2 a PhD), while 9 had finished their education during high-school.

Experiment 2, due to its on-site nature, had a smaller sample. A total of 26 subjects (19 men and 7 women) participated, and were randomly distributed in groups of 13 before taking each version (2A and 2B) of the test. Average ages were 25 years old in group 2A, and 24.31 years old in group 2B. The first group had people of ages ranging from 18 to 38, while in the second group went from 18 to 37. 12 participants were studying a degree related to Computer Science, 9 of them had already obtained it, 4 had a PhD in this same field and 1 had finished a master's degree.

5.5 Results

5.5.1 Experiment 1

The aim of the first experiment was to analyse differences in accuracy when users try to locate a sound coming from behind, and general performance while doing so with and without an audio spatialisation system based on LPF. As only one sound came from the rear in each case, accuracy was measured through the amount of right answers for that particular sound.

TABLE 5.1. Accuracy (right and wrong answers) when identifying sound direction for groups 1*A* and 1*B*.

Group 1A	**First sound**		**Second sound**		**Third sound (rear)**		**Fourth sound**	
	Right	Wrong	Right	Wrong	Right	Wrong	Right	Wrong
	8	21	7	22	9	20	3	26
Group 1B	**First sound**		**Second sound (rear)**		**Third sound**		**Fourth sound**	
	Right	Wrong	Right	Wrong	Right	Wrong	Right	Wrong
	13	16	12	17	4	25	4	25

These first results meant there were high rates of failure among most subjects, and thus could not be considered statistically significant. Table 5.1 shows a success rate (number of right answers divided by the total amount of subjects) for group 1A of only 31.03 %, and group 1B achieved a 41.38 %. Even though the difference was of more than 10 points, not having any set of answers with a success rate of more than 50 % pointed to the conclusion that the experiment was overly complex for a person with normal hearing and no specific education in music or sound design. These results, however, were promising, as group 1B achieved better results and was making use of LPFs.

5.5.2 Experiment 2

Experiment 2 was designed with these difficulties in mind, and aimed to be accessible to everyone. Its results were more enlightening, as can be seen in Table 5.2 and Figure 5.4. Average time taken by subjects from group 2A to complete the task was 38.84 seconds, and participants from group 2B took 23.66 seconds on average to complete the same challenge. Due to the only difference being the new, LPF-based spatialisation system in group 2B, it is possible to assert these results were caused by it.

TABLE 5.2. t-Student test for times achieved by users in $2A$ and $2B$.

Group	**2A**	**2B**
Subjects (N)	13	13
Average	38.8431	23.6600
Std. Deviation	19.1029	7.1668
SEM	5.2982	1.9877
P**-value**	0.0130	

Results from this experiment were not regular, however, mostly due to the variation induced by the different levels of ability found in subjects when locating sounds in a virtual environment. However, maximum and average completion times follow a similar trend: 2A hit a maximum of 80.36 seconds, whereas 2B's highest time was 41.4 seconds.

Subjects were asked about their own performance after completing the experience, with the question: "Was it easy for you to identify sound positions during the experiment?", and their answers were the following:

11 people in group 2B answered in a positive way: "Strongly agree" (3) or "Agree" (8), and 2 gave neutral answers ("Neither agree nor disagree"). Subjects in group 2A, however, gave 8 positive answers: "Strongly agree" (3) and "Agree" (5), and 5 neutral answers.

Therefore, results achieved with the LPF-based system negate the null hypothesis (H_0) and confirm the alternative (H_1). There existed a notorious, statistically significant difference in performance between both systems (2A and 2B), and a regular distribution of subject auditory skills is assumed in both groups. That leaves the addition of LPF sounds coming from behind as the only controlled variable responsible for the variation in performance or accuracy between groups.

Women were distributed evenly between groups during both experiments, but their small numbers could have influenced the results, as differences in hearing between men and women have been previously discovered [19].

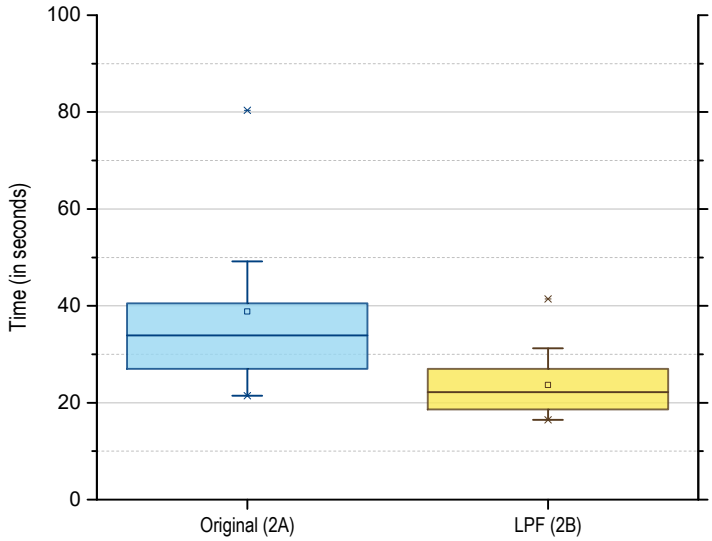


FIGURE 5.4. Comparison between average times achieved by users in both the original (2A) and improved with an LPF (2B) sound systems.

5.6 Conclusions

Due to the fact that A and B tests contained a single difference in sound processing, and considering the results are better in B groups for both experiment 1 and experiment 2, it is possible to reach the conclusion that the addition of LPFs to rear sounds seems to improve recognition of sound location in the virtual environments tested. The implementation of a LPF-based system does not increase realism (it reduces it instead), but identification accuracy improves. As LPFs are only applied when a certain sound comes from an off-camera place, most players react by rotating when this effect is detected. The learning curve in this process is quick, and happens as soon as a user has checked the position of a sound with this effect applied.

Participant's perception of self-performance was very high in both groups during experiment 2. Even subjects with the highest time scores gave neutral answers to the question on this matter, which leads to the conclusion that there was no conscious advantage detected for subjects in group 2B; their results were, however, much better.

Consequently, designing game audio with recognition in mind (instead of pure realism) can prove useful in certain situations. Sound design strategies like the ones presented here can contribute to a better and quicker location of in-game objects, without the need to implement more complex, HRTF-based audio.

Findings presented in this chapter opened the door for further experimentation with 3D positional audio and the conception of a separate architecture for adaptive music generation in real time. This new system is described in Chapters 6 to 8.

CHAPTER 6

LitSens: An emotion-driven adaptive music system

"I don't want to be at the mercy of my emotions. I want to use them, to enjoy them, and to dominate them."

Oscar Wilde

As was explained during Chapters 2 and 4, emotions can have a big impact over how players behave in an interactive experience. Sound, and especially music, have the power to alter these emotions, and thus can also become behavioural modifiers.

As part of this research, an adaptive music system based on emotions was designed, with the intention of improving behavioural game design. This system is called LitSens, and its functioning will be detailed in the following sections.

6.1 LitSens' initial architecture

LitSens is an adaptive music system which is designed to create atmospheres that match emotions arisen during a video game experience. This system does not generate music from scratch, due to the limitations found in procedural techniques and explained in Chapter 2; instead, it comes up with combinations of pre-recorded fragments of music to create unique soundtracks for different situations.

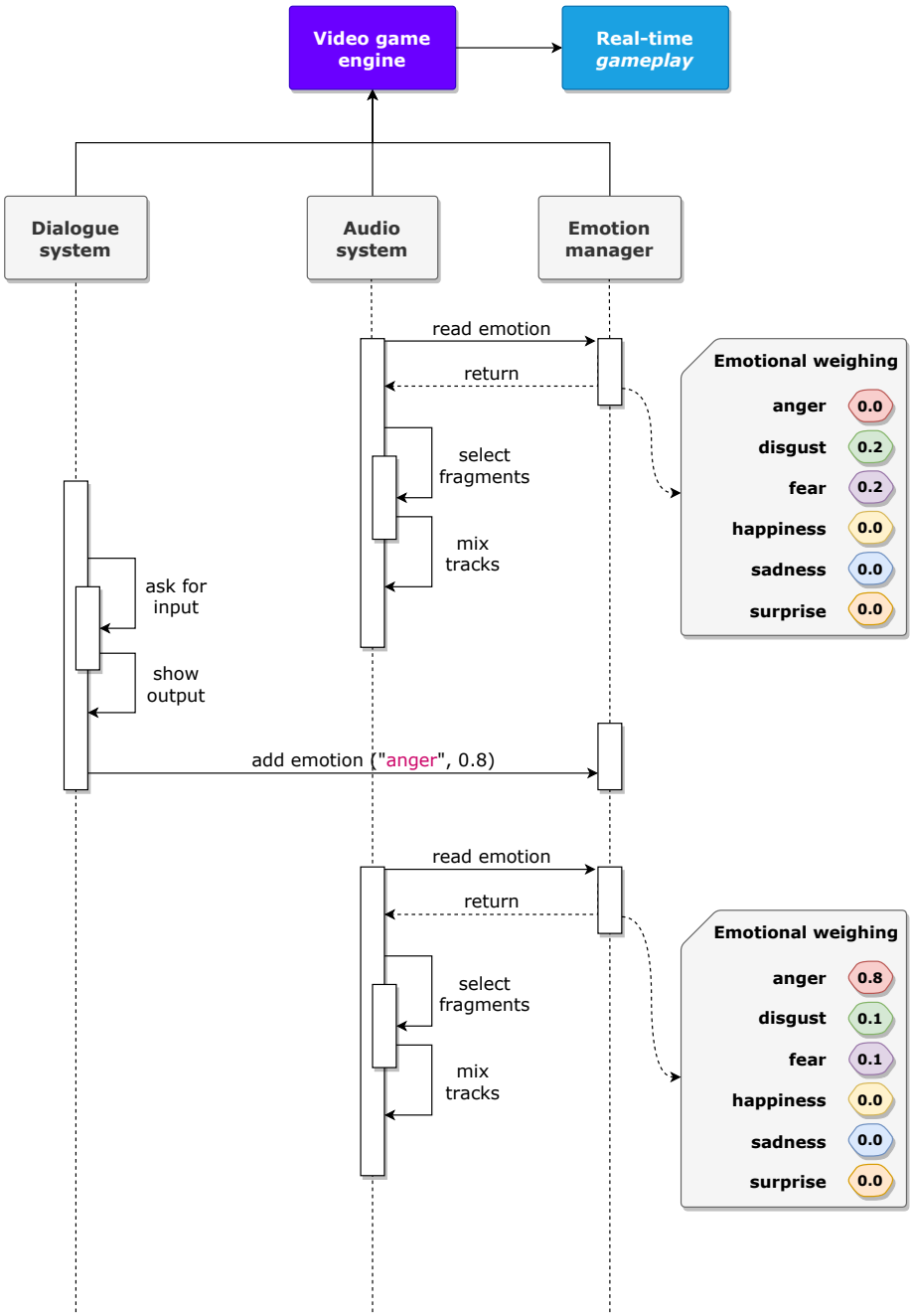


FIGURE 6.1. Graphical representation of the data flow in LitSens v0.1. The dialogue system triggers interactions by asking for input and showing an output sentence, tagged with an emotion value. The emotion manager stores emotion data and changes the value of each variable over time. The audio system reads emotions from the emotion manager, selects a series of relevant musical fragments, adds them to the mix through the audio engine and repeats this process until the experience ends.

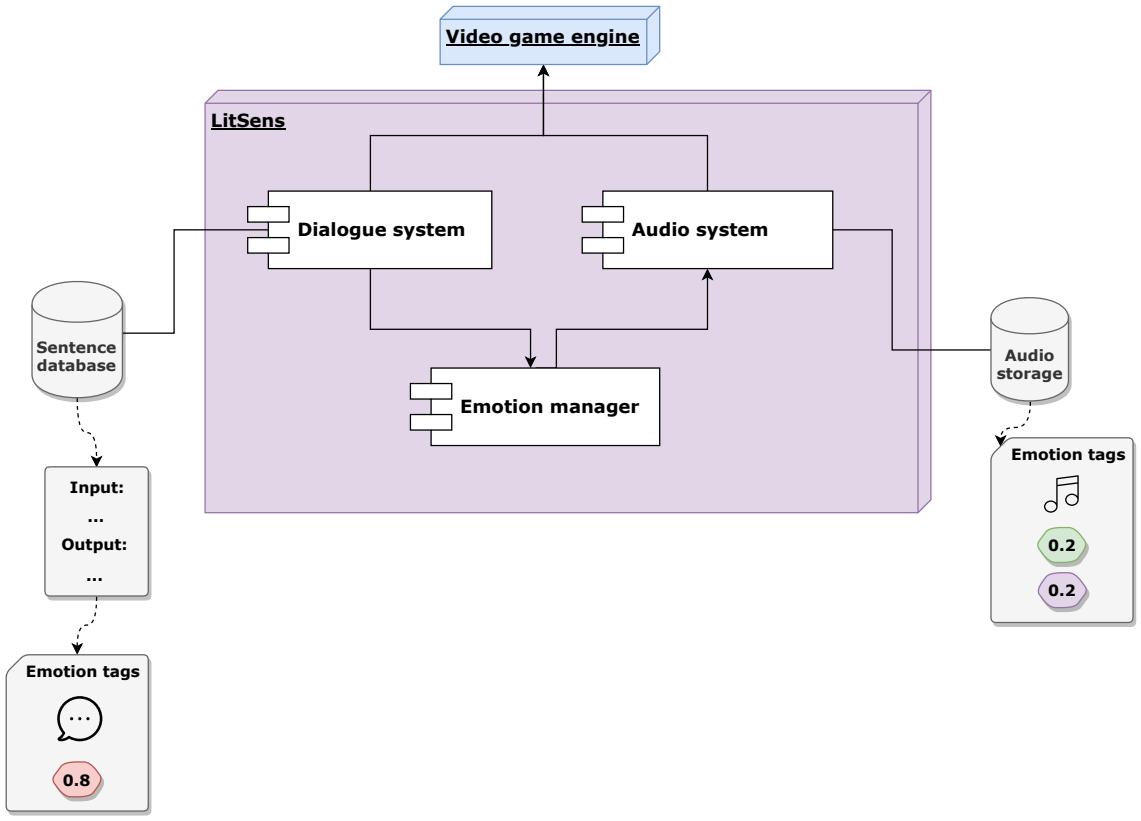


FIGURE 6.2. System diagram of LitSens v0.1. The initial architecture includes three modules: a dialogue system, an audio system and an emotion manager, which interact as described in Figure 6.1. The dialogue system reads from a sentence database and offers input and output options to the player, all of them tagged with at least one particular emotion. The audio system reads from a fragment database, in which every item is also tagged with at least one emotion.

In its original form, LitSens' architecture followed the pattern depicted in Figures 6.1 and 6.2, and worked as follows [69]:

– The game includes a dialogue system which the user can interact with in two ways: receiving messages from a non-player character (NPC) (*output sentences*) and choosing answers to those messages (*input sentences*) from a series of options.

– The *dialogue system* takes information from a database of tree-form short dialogues (*stored sentences*). These dialogues were designed by a human, and try to induce a variety of emotional responses in the player. In the initial setup, players can choose from 3 different answers each time a text output is shown on screen, and this act is understood as the expression of a feeling.

- Every possible answer has an "emotional weighing", containing a value in a range from 0 to 1 for one of the 6 basic emotions by P. Ekman [24]. For example: receiving a declaration of war would have an anger value of 1, and watching a butcher work would have a disgust value of 0.5.

- The emotional weights of caracters' messages and player choices are stored, so that the system can monitor the emotional context of a playing session. With each addition of an emotion, the state of the *emotional weight storage* changes. A maximum of 3 concurrent emotions are selected at the same time, with an intensity that depends on their accumulated weight. For example, the appearance of three consecutive emotions with the following values: `[anger = 1; sadness = 0.3; sadness = 0.4]` generates the following state: `[anger = 1; sadness = 0.7]`.

- A maximum of 3 tracks (adapting to 3 different emotions) is established for several reasons. On one hand, emotions can overlap (e. g.: "anger" and "disgust" generate "loathing"), which means at least 2 simultaneous tracks should be present to recreate this complex atmosphere. Besides, a third track is useful when inducing remembrance in players. If, for example, "anger", "disgust" and "happiness" are concurrent emotions, it would be possible to combine 3 tracks to create a sense of loathing while hinting at a lost happiness.

- Because of this, the 3-track audio system decides which fragments of music are to be played each delta second. Once in that period of time, stored emotion variables are read and a new piece of music related to those feelings starts playing if necessary. If no additional emotional responses appear, the previous soundtrack will keep playing in loop until the application closes.

- All selected tracks are chosen from a database of 60 pre-designed fragments of music, created by a human composer, sharing a tempo of 110 beats per minute. All of them are seamless loops, with lengths that range from 5 seconds to 15. A link to an example of one of these fragments can be found at the bottom of this page[1]. Therefore, this system is not designed to allow for quick sequences of changes in ambient music (of less than 5 seconds), as some tracks could be abruptly interrupted.

- Every time a new track is selected and played, it adjusts to a fixed "tempo grid", skipping up to a beat if necessary. Once a combination of two or more tracks is playing, they are mixed in the audio engine. Intensity (volume) is established depending on the values of emotional weight, for a normalised intensity of 1. Primary emotions have a normalised (non-logarithmic) intensity of 0.5, followed by secondary (0.3) and tertiary emotions (0.2).

[1] `https://soundcloud.com/asomnu/litsens-short-fragment-sample/s-FA2gy`

– The resulting musical atmosphere plays through the game engine's audio system. This makes it possible to spatialise or add extra layers of effects to each track. At the bottom of the page[2], a link to an atmospheric track generated using the default's LitSens database, based on the emotion P. Ekman calls "fear", is included.

The system described above aims to constitute a versatile option for having emotion-driven ambient music without the low quality sound and high processing cost of a fully procedural architecture. Though the design of this technology was initially aimed for Unity 3D, it is potentially compatible with any game engine that allows for multi-track mixing in real time.

Moreover, the audio managing technique this system utilises acts more as an "assistant" than a "proxy" for musicians. As a consequence, the final result can look exactly like a human-made composition, but it is shaped by both human creativity and computational adaptability. This means the resulting products are soundtracks which are more flexible, but that do not lose the "touch" of a human composer's hand. Expert knowledge is needed for the system, but musicians that use it are given the capacity to recreate realistic music tracks for video games or interactive experiences, while maintaining instantaneous adaptability.

6.2 A new development using sentiment analysis

The next iteration for LitSens' architecture was to aim for removing the need to manually tag every emotion involved in the selection process. Considering a text-based approach had been taken during the initial design phase, sentiment analysis proved to be a useful addition in this case scenario.

As can be appreciated in Figures 6.3 and 6.4, some changes were introduced to the initial architecture, and the flow of events during a standard experience had been slightly modified so as to make it possible to test the new features. The renewed system worked as follows [68]:

– This conception of LitSens constitutes an interactive experience in which there is narrative content in the form of text.

– For the design of this system, three main components are taken into account: a dialogue system, an audio system and an emotion manager.

– At the beginning of the experience, a text with narrative content is shown in a text box. Users can then use a text field to input a response, which will produce an output consisting of the next sentence of the narrative.

[2] `https://soundcloud.com/asomnu/litsens-horror-track-sample/s-0ezcq`

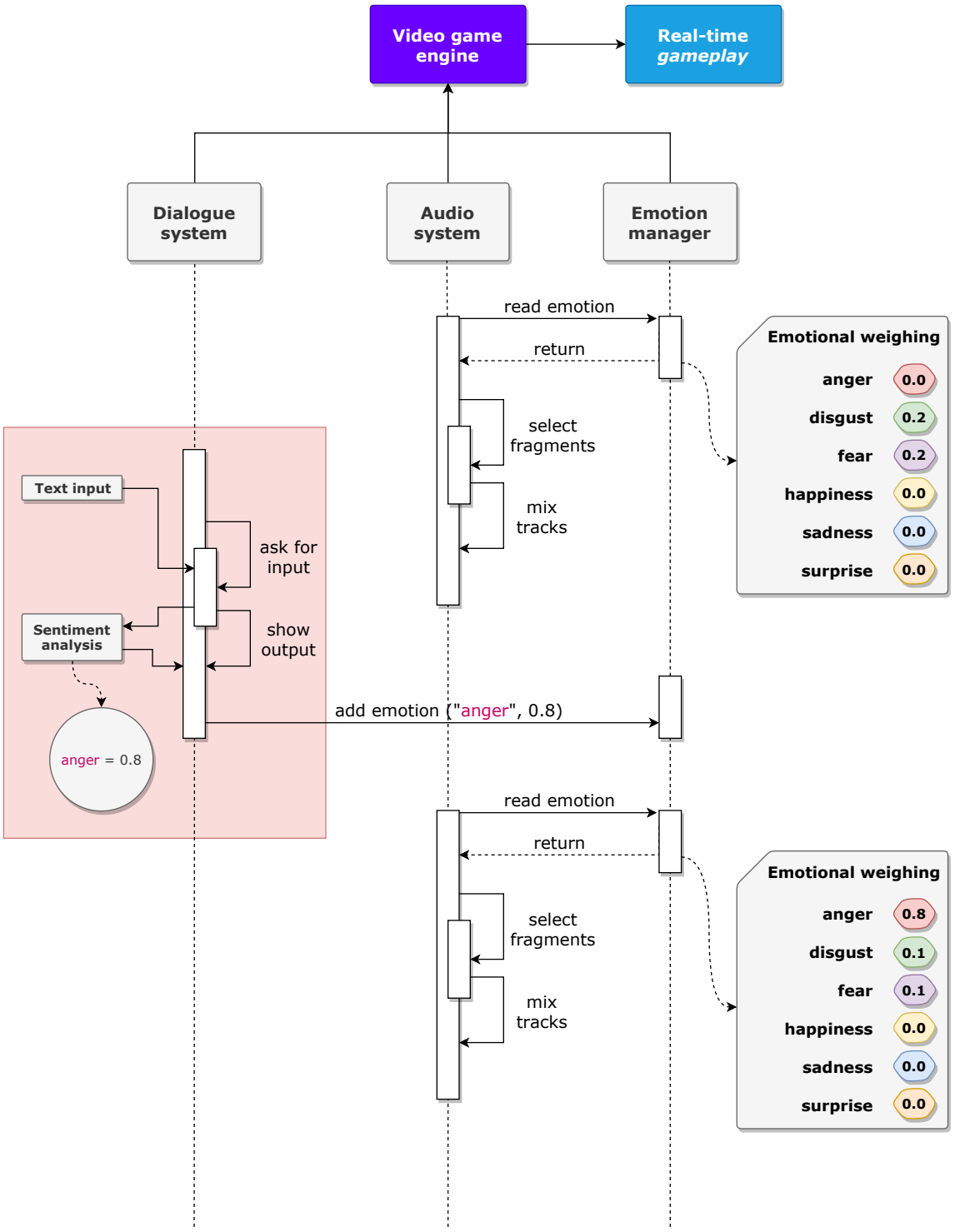


FIGURE 6.3. Data flow diagram of the second iteration of LitSens (v0.2). Text input and sentiment analysis (marked by a red background) are added.

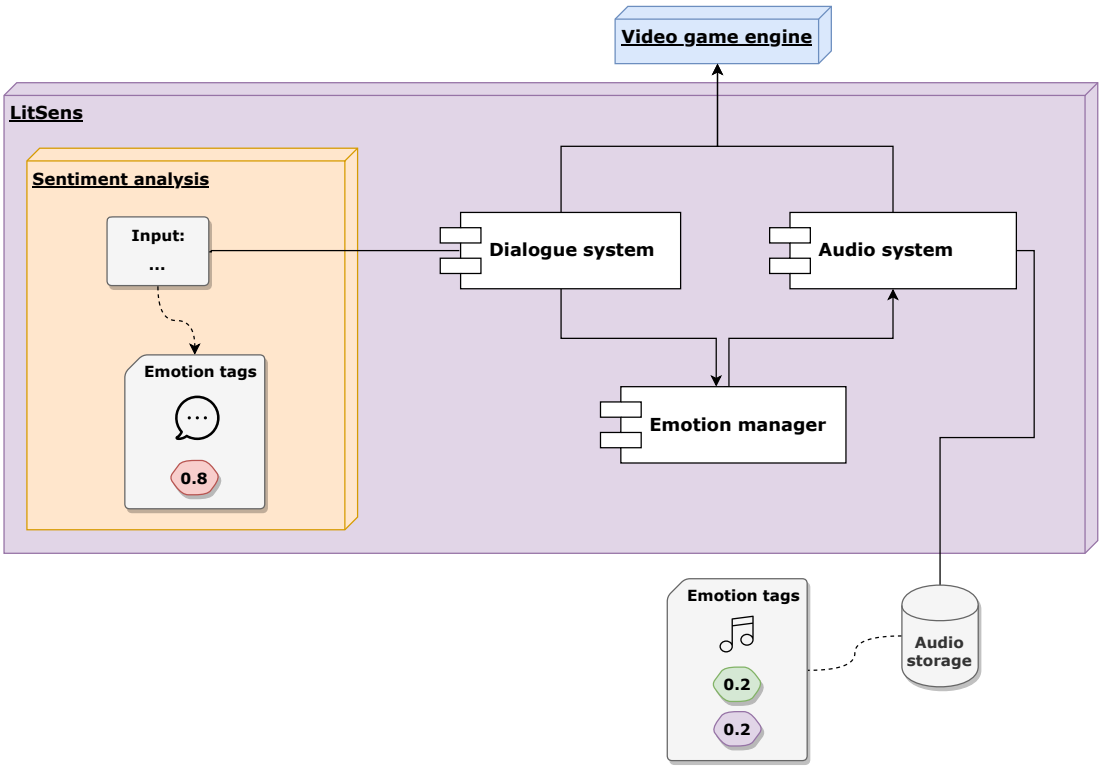


FIGURE 6.4. System diagram of LitSens v0.2. The design now includes an additional system for sentiment analysis, which interacts with the game engine through the dialogue system.

– The input is then processed by a sentiment analyser, which provides several values. The three values with more weight are then normalised and stored by the Emotion Manager.

– The audio system reads those values from the emotion manager, and selects a maximum of three music fragments, which start playing through the game engine after a simple mix process. As shown in Figure 6.4, all music fragments are previously designed by a human, tagged with an emotion and stored in a database.

– Following inputs will modify the state of the emotion manager, producing different fragment combinations and causing them to change in real time.

– A total of 6 emotions are considered, following Ekman's taxonomy of basic emotions [24]: happiness, sadness, fear, surprise, disgust and anger.

– A 3-layer system was chosen so as to be able to represent multiple emotions at the same time. As Jørgensen [49] states (see Chapter 2), game audio has a dual role: it supports the general feeling of an environment, but also gives vital information during *gameplay*. By having three layers, it is possible to include an atmospheric track based on past

feelings, as well as common dual emotions like "happiness-sadness", "fear-disgust" and "disgust-anger".

This step in the development of LitSens worked as an introduction for later iterations, where more complex emotion recognition was implemented. The implementation of a sentiment analyser was discarded, eventually, in favour of different approaches that will be described from now on, and which adapted better to first-person and VR environments.

CHAPTER 7

INFLUENCING PLAYER BEHAVIOUR WITH LITSENS

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."

Douglas Adams

Experiments described in Chapter 4 hinted at a correlation between variations in the basic elements of a certain soundtrack and player decisions during an interactive experience: harmonic, high-pitched melodies seemed to attract users more efficiently than cacophonous, low-pitched ones. Moreover, different behavioural tendencies were observed in different users, depending on their results when taking the Bartle test [6]; some groups of subjects were attracted to musical attributes that did not work as a lure for different profiles. Thus, personal auditory preferences seemed important when using audio to create certain types of behaviours in an interactive experience.

Previous research with blind people by O. Lahav [56], as stated in [70], "acknowledges the existence of a conceptual level, in addition to a perceptual one, in the learning process associated with scouring an unknown virtual environment in search for clues that allow to build mental 3D maps". Lahav's point of view is very important for this research, as it upholds the existence of mental states and categories, which are attained through education and culture, in strict relationship with formal auditory parameters, such as pitch or intensity.

The next step in the development of LitSens was to make it work as a behavioural design

tool, now working on Unreal Engine 4, which does not rely on GUI elements or textual narration to induce certain player actions. The following sections explain how the system was modified and the experimental phase which took place in order to validate its properties.

7.1 A new version of LitSens

During this phase, LitSens was modified so that it did not just recombine musical fragments: it was also able to alter three different audio properties in real time. These properties were: musical complexity, pitch and rhythm.

Complexity was increased by introducing layers of sound which could disrupt the harmony of a base track. This alteration can be appreciated when comparing the two spectrograms in Figure 7.1.

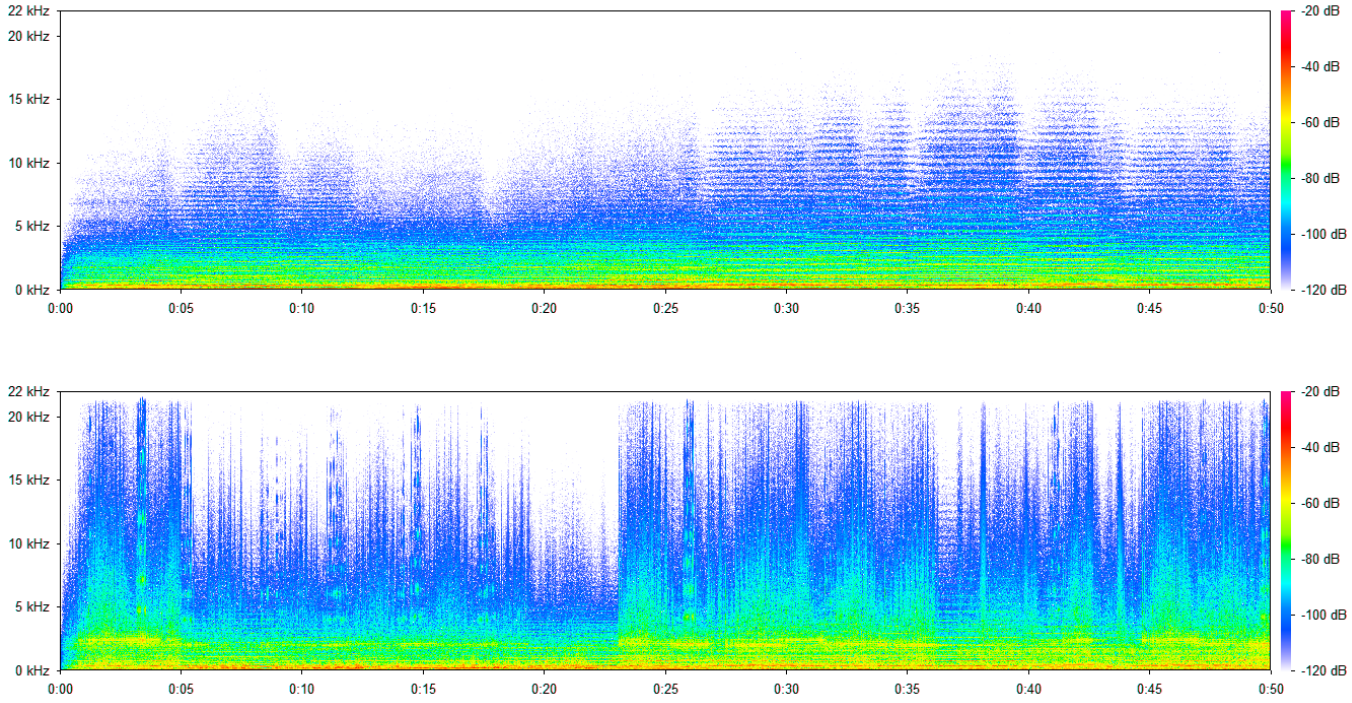


FIGURE 7.1. Spectrograms depicting simple (top) and complex (bottom) variations of the same sound.

Pitch was modified by the game engine (Unreal Engine 4) in real time, while maintaining a maximum variation of a 50 %, to avoid unwanted artefacts. The effect was applied to whole fragments of audio, and always started and ended with them.

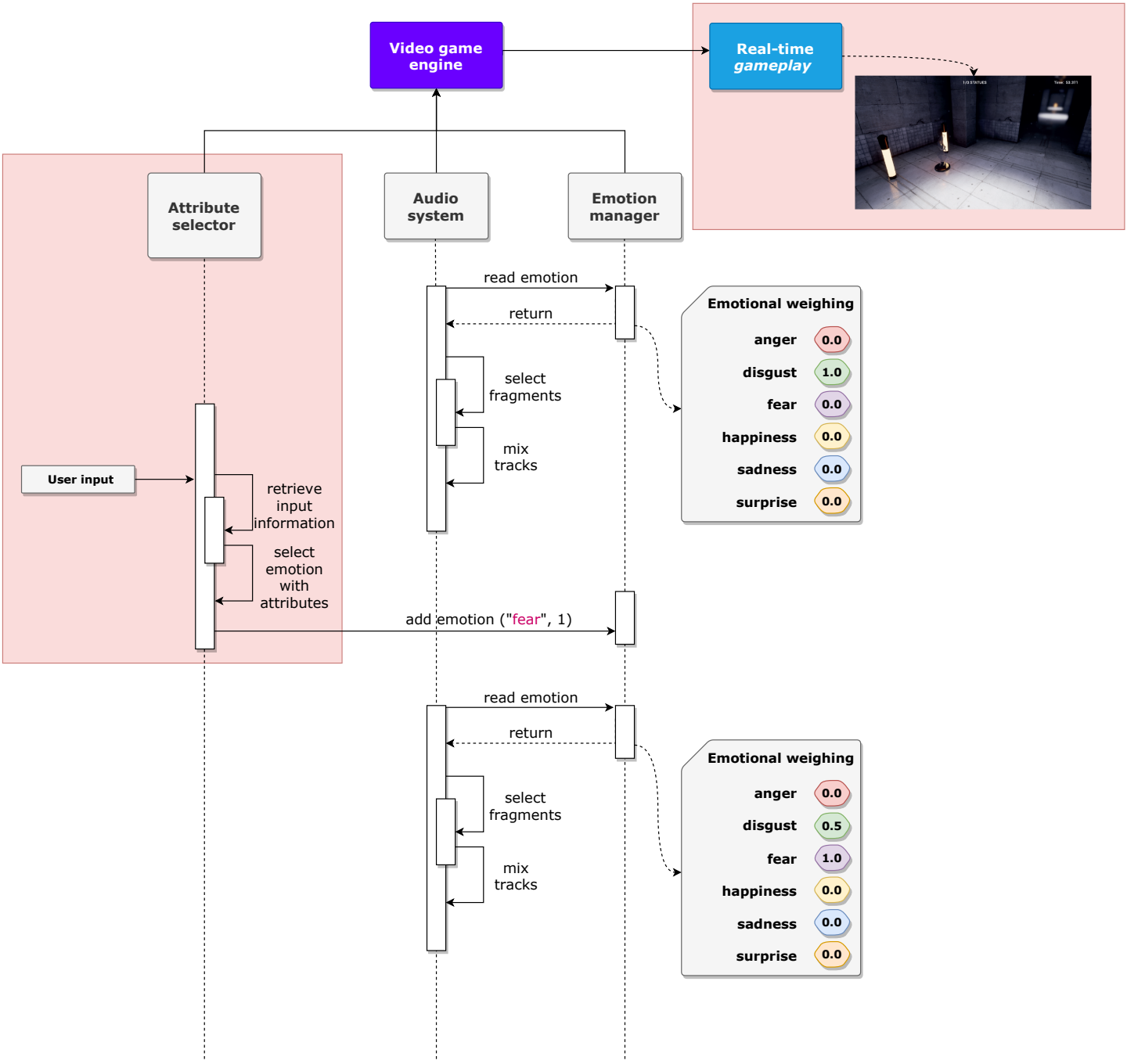


FIGURE 7.2. Data flow diagram for LitSens v0.3. Users were now able to select specific musical attributes when playing a prototype in Unreal Engine 4. A red background is included in certain areas to show all the changes made from the previous version.

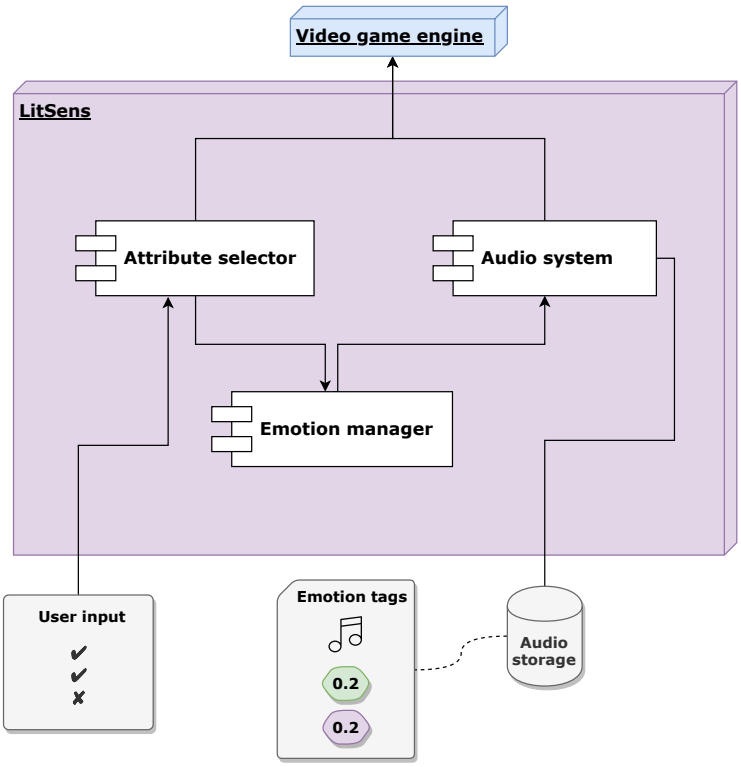


FIGURE 7.3. System diagram of LitSens v0.3. A new attribute selector is added so that players can choose from three musical parameters.

Rhythm was also changed using built-in game engine features, and was applied in the same way as pitch variations (it was fragment-dependant). A variation limit of 25 % of the original tempo was applied to reduce the amount of distortion and timbre loss which usually happens when applying this kind of effect. All these changes can be appreciated in Figures 7.2 and 7.3.

7.2 Experiment design

Aiming to test the system described above, and to explore the relationship between adaptive music and player behaviour, a new experiment was designed. The experience consisted of a labyrinth-like orientation puzzle which players had to solve with or without adaptive music; the level of consistency between users' perception of sounds and their actual performance and response to them was also measured, as is detailed in [70].

7.2.1 Design

When the experiment began, all participants were randomly distributed in two groups, which from now on will be called A and B. Initially, both groups had the same size (N = 17), though group A lost a subject due to hearing health problems. Throughout the experiment, only two persons were in the room at a time: one participant and one test supervisor. Four different phases took place in each session: Self-Assessment Manikin (SAM) test [13, 34] (see Chapter 2), attribute selection, *gameplay* and sociological survey, and followed this order.

Participants in group A started by taking a SAM test (detailed in Appendix D) about three pairs of sounds. Each pair was played consecutively, and had a strong relationship with one of the basic categories used to classify sounds in the test-bed game that will be described later. Sounds in every pair were intended to represent the two opposing concepts in each of the following categories, presented in order during the test: tone (low-high), structure (simple-complex) and rhythm (slow-fast). The differences between sounds in each category were big enough to be easily noticeable, as can be appreciated in Figure 7.1. Besides, during this test, all sounds were evaluated separately after listening to each pair, in order to compare them.

The SAM test was passed in its 9-point scale version, by means of a digital form which contained all three measurements: emotional valence, arousal and dominance. This test uses the Semantic Differential [74] as a basis, and simplifies it. Thus, *emotional valence* measures "pleasure", and is strongly related to bipolar adjective pairs such as "unhappy-happy", "annoyed-pleased", "unsatisfied-satisfied", "melancholic-contented", "despairing-hopeful" or "bored-relaxed". *Arousal*, on the other hand, is related to pairs like "relaxed-stimulated", "calm-excited", "sluggish-frenzied", "dull-jittery", "sleepy-wide awake" and "unaroused-aroused". Lastly, *dominance* is associated with adjective pairs like "controlled-controlling", "influenced-influential", "cared for-in control", "awed-important", "submissive-dominant" or "guided-autonomous".

However, subjects from group B were given only one sound to evaluate, in spite of taking the same kind of test. Said sound contained the default audio played by their version of the game, classified as: slow, low and simple. This evaluation was not taken into account later and it was performed to give the subject of this group the same insight than the subjects in group A about the auditory nature of the experiment, in order to avoid possible bias.

Once the SAM test had been passed, subjects from both groups had to launch a PC video game made expressly for this purpose, which ran at 60 frames per second in a 24-inch LCD screen and was controlled by a mouse and a keyboard.

Participants in group A (considered the experimental group during this test) were asked to select, from an in-game menu with three categories ("rhythm", "tone" and "structure"), the attribute for each of them ("slow-fast", "low-high" or "simple-complex"), which, in their opinion, would make a sound stand out over the rest. Consequently, a total of 8 outcomes were

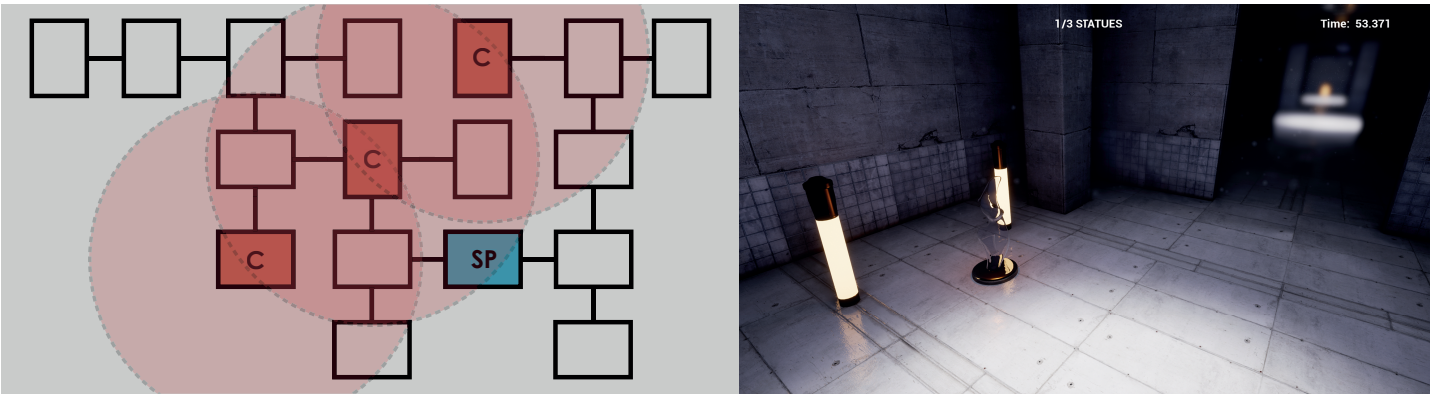


FIGURE 7.4. Diagram showing the layout of the virtual environment utilised during the experiment. There is a starting point (SP) and three collectables (C) in the form of small statuettes. Red circles represent the area in which each sound could be listened to. Walls applied occlusion through a low-pass filter, not depicted in this diagram. A capture of the game from a first-person perspective is also included on the right.

possible, following this pattern: "X rhythm, Y tone and Z structure".

Group B was considered the control group in this context, and participants in it were not given the option to choose their own audio configuration; instead, they played with a default audio track (with the parameters "slow rhythm", "low tone" and "simple structure" selected).

No sound clues were included in the game's menu to help users from group A decide: the only previous sound reference given was the SAM test. This decision was made so as to be able to evaluate the coherence between subjects' perception of what sound suits them better in a certain situation and actual performance produced by their selection.

After people in group A selected a combination of parameters, a personalised level was loaded. For group B, on the other hand, the level loaded with the default audio track. In both cases, said level consisted of a three-dimensional labyrinth, which could be navigated from a first-person perspective. From a strictly logical standpoint, however, its structure can be defined as two-dimensional, as there were no important interactions in the vertical axis. This labyrinth is depicted as a map in Figure 7.4.

Players could move around and look using the specified control system: a keyboard (WASD keys) and a mouse. Each participant was told to look for and recover a total of three small figures or statuettes inside this labyrinth, as quickly as possible. Elapsed time and number of statuettes recovered were shown on the screen permanently to keep players informed at any time about their goal. The only possible method for recovering a statuette was to step on it —namely: to overlap its collider with the player's. Every time one of these statuettes was picked up, a measure of total elapsed time was stored in a log file. At the end of each

session, this log was automatically retrieved and tagged with the correspondent participant number. Thus, three time measurements were taken for each subject. For convenience, these measurements will henceforth be called "t_1" (first statuette), "t_2" (second statuette) and "t_3" (third statuette, or total time elapsed).

Every one of these three figures emitted a spatialised, monophonic music track which blended with a base stereophonic soundtrack. Said soundtrack was a low, synthetic drone, with no variations in tone or intensity. For users in group A, the spatialised track was modified to adapt to their specified preferences in musical attributes, as is described above. People in group B, however, had a default track playing from every statuette. In any version of the game, the statuettes stopped emitting sound when they were recovered, by means of a 2 second linear fade out. If the spatialised audio track was received by the camera listener through a wall, a low-pass filter with a cutoff frequency of 900 Hz was applied.

When the three figures were recovered, the game automatically ended and the application was closed. Once the game had been played, every subject had to take a brief survey to determine their sociological profile. Data retrieved in this test included: age, sex, country of birth, level of education completed, presence of hearing problems, fondness for music and sound and performance when playing video games. Participants were also asked, after completing this survey, if sound was useful when trying to find the three objects inside the virtual labyrinth. Results from this question were stored in a variable that was called "help index" (h_i). A Likert 5-point scale [59, 84] was utilised for this and all questions requiring gradation, except for the SAM test, where the standardised 9-point scale was employed.

Additionally, two leaflets with instructions were created: one for group A and one for group B. Every participant had to read only the pertinent one while waiting for the experiment to begin. These documents contained a detailed description of all actions every user would have to take during the experiment. Brief instructions on how to listen to the sounds and how to take the SAM test were included, as well as the keyboard and mouse controls for the video game. All subjects were also told it was of utmost importance to complete the level in a time as short as possible, and that to do so they would have to find three small statuettes. The only difference between "A" and "B" versions was the lack of explanation on how to evaluate pairs of sounds (since this was not necessary for group B).

A parametric, unpaired test (Student's t test) [40] was utilised to evaluate results from both the 5 point and 9 point scales, as well as for times (t_1, t_2, t_3).

Finally, after subjects from group A completed the experiment, they were asked to explain, in their own words, the reasons for their attributes selection, and notes were taken.

7.2.2 Hypothesis

The main hypothesis behind this experiment was that a statistical difference may be found between the two groups of subjects (A and B), in terms of performance (measured in total time or t_3), in the virtual environment described above. The independent variable is the presence or absence of a preference selector at the beginning of the experiment that influences music played during the game. It was also intended to find a relationship between the initial selection of auditory features (available to participants in group A only) and t_3.

7.2.3 Demography

Participants had to meet at least two prerequisites so as to be able to take the experiment: they had to be able to hear properly, and they had to be familiar with at least one video game of the first-person shooter (FPS) genre.

The experiment took place in a university in Spain, and all subjects were students (graduate and postgraduate) or worked as lecturers in the Computer Science field.

A total of 33 subjects participated in the experiment, of which 16 were assigned to group A (composed of 14 males and 2 females) and 17 to group B (with 15 males and 2 females). 29 of these people were born in Spain, and the remaining 4 were born in Colombia, Bolivia, Switzerland and Venezuela. All of them were native speakers of Spanish, and this language was used throughout the whole experiment; materials quoted here have been translated into English for convenience. Besides, all participants shared similar cultural features, and all but one had lived most of their lives in Spain.

Average ages in groups A and B were very similar: 23.438 for group A and 24.059 for group B. The mode was 18 in both cases, as most participants were freshmen.

68.8 % of the participants were undergraduate students, whereas 6.3 % were studying a master's degree at the moment. The rest were Ph. D. students (12.4 %) or university professors and researchers (12.5 %).

When asked if they played games frequently, most subjects in groups A and B answered positively, with a mode of 5 out of 5 in both cases, and a mean of 4.5 (for group A) and 4.412 (for group B). Most participants also considered themselves good video game players, with a mode of 4 out of 5 for both groups and an average of 3.688 (A) and 3.824 (B). The scores were slightly lower when asking them about their perceived performance in FPS games: the mode was 3 out of 5 in A and B, while the averages were 3.5 (A) and 3.353 (B).

As for self-evaluation of hearing proficiency, when asked if they had good hearing, the modes were 5 out of 5 in group A and 4 out of 5 in group B; the averages were 4.125 (A) and 4 (B). Moreover, when told to answer if they had a "good ear" for music, the mode was 4 out of 5 in both cases, and the averages were 3.688 (A) and 3.353 (B).

There were 4 exceptions to the random distribution of subjects between groups: musicians were selected and distributed evenly, with a total of 2 of them in each group. This was done to avoid possible bias due to their knowledge of music and audio, and they were the only participants which were not randomly distributed. This process took place before starting with the experiment, and the affected participants were unaware of it.

In conclusion, the surveyed sample had a very good perception of their own hearing, but their confidence in their musical ear was average-to-neutral. They were also frequent *gamers* and had a good knowledge of the media, and most of them had technical knowledge about how video games work or how they are made.

7.3 Results

Results from the described experiment and its associated survey pointed to several statistically significant differences between groups A and B in terms of both performance and self-assessment.

Participants in group A achieved a total average time of completion (total time or t_3) of 78.108 seconds, whereas participants in group B took an average of 132.987 seconds to complete the same puzzle. The median in group A is 75.694, while in group B is 100.668 seconds. The lack of similarity between average and median times in group B can be explained by the presence of two clear outliers (as can be appreciated in Figure 7.5). These two persons completed the game in 369.250 and 367.020 seconds respectively. A parametric analysis of these results can be found in Table 7.1. After a Student's t-Test, the two-tailed P value is 0.0356, which makes differences in these datasets significant from a statistical point of view.

Due to the fact that total time was measured when every statuette was picked up, not only total elapsed time gave an important insight about player behaviour during the experiment. It is also quite illustrative to look at how the difference in average time between the two groups increases as every object is taken (see Figure 7.6). t_1 had an average value of 22.068 for group A, and of 25.637 for group B, so the difference between means equals a mere 3.569 seconds. On the other hand, t_2 had an average value of 41.854 for group A and of 53.398 for group B, producing a difference of 11.544 seconds. Lastly, t_3 presents the biggest difference: 54.879 seconds.

As for the help index (h_i), it also presents statistically significant differences between groups. As can be appreciated in Table 7.2, group A has a mean of 3.56 out of 5, while group B scores 2.47 points. The mode is especially enlightening in this case: 5 in group A and 1 in group B. After a variance analysis (again, a Student's t-Test), the P value is 0.0484 when comparing the two datasets.

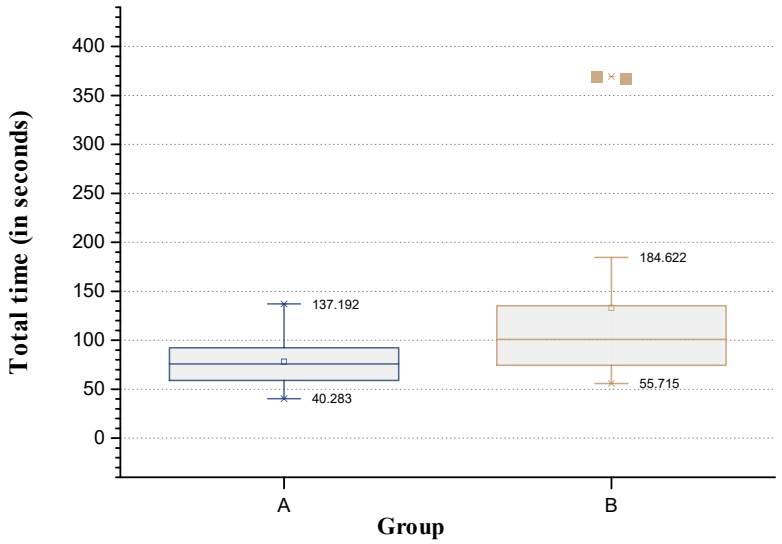


FIGURE 7.5. Differences in performance between groups A and B, measured in total time of completion (t_3).

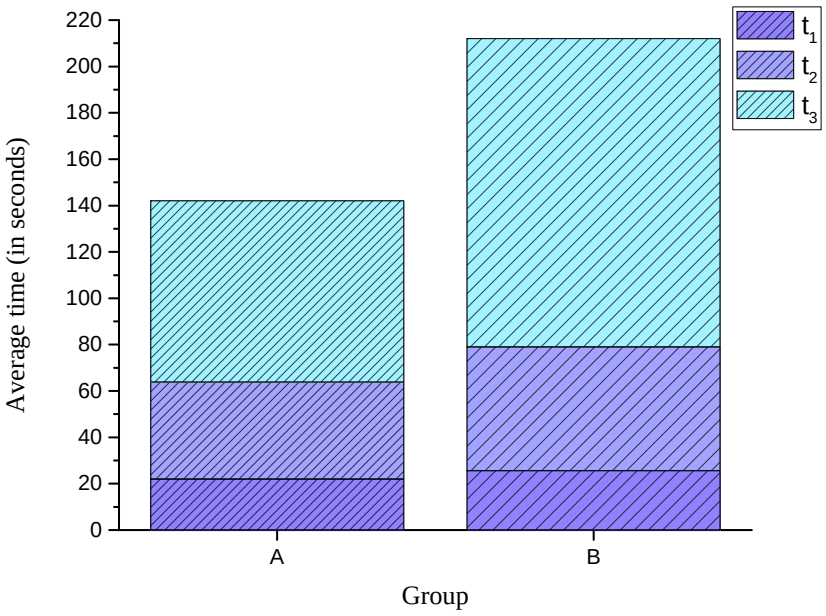


FIGURE 7.6. t_n increments for groups A and B.

TABLE 7.1. *Student's t-Test for total time (t_3) in groups A and B.*

Group	**A**	**B**
N	16	17
Mean	78.108	132.987
Standard deviation	27.908	96.090
Two-tailed P value:	0.0356	

TABLE 7.2. *Student's t-Test for h_i values in groups A and B.*

Group	**A**	**B**
N	16	17
Mean	3.56	2.47
Mode	5	1
Standard deviation	1.46	1.59
Two-tailed P value:	0.0484	

There does not exist a significant statistical relationship between t_n and the initial selection of auditory features, which was only possible for members inside group A, as Table 7.3 shows. "Simple" (11), "high" (9) and "fast" (9) were the most common options, in spite of everything mentioned above.

TABLE 7.3. *Features selected by group A participants and total time achieved (t_3).*

t_3	**Tone**	**Rhythm**	**Complexity**
40,283	High	Fast	Simple
42,159	High	Fast	Simple
45,318	High	Fast	Simple
57,027	Low	Slow	Simple
60,738	Low	Fast	Simple
66,320	Low	Slow	Simple
73,127	Low	Slow	Complex
73,483	High	Slow	Simple
77,904	High	Fast	Simple
81,639	High	Fast	Simple
84,315	High	Fast	Complex
92,305	Low	Slow	Complex
92,315	Low	Slow	Simple
95,468	Low	Fast	Complex
130,129	High	Fast	Complex
137,192	High	Slow	Simple

It is also worth noting that the attribute "complex" was the least selected, as only 5 subjects chose it. However, this same attribute achieved the highest dominance score during the SAM test, with an average of 5.875 and a mode of 7 out of 9. Not only that, but it also had the highest excitement score, averaging 5.688 and with a mode of 7 out of 9. Additionally,

general results from the SAM test were not consistent with player selections of attributes before playing the game. This can be appreciated in Table 7.4.

TABLE 7.4. *SAM test results in 9 point scale for variations of the same sound.*

Attribute	SAM scale	Average	Mode
1. High tone	Valence	5.938	7
	Arousal	3.625	2
	Dominance	4.5	5
2. Low tone	Valence	4.697	3
	Arousal	3.152	2
	Dominance	4.727	3
3. Simple structure	Valence	5.688	4
	Arousal	3.563	3
	Dominance	4.188	3
4. Complex structure	Valence	3.375	5
	Arousal	5.688	7
	Dominance	5.875	7
5. Fast tempo	Valence	6.063	7
	Arousal	5.25	7
	Dominance	5.188	5
6. Slow tempo	Valence	5.375	6
	Arousal	3.438	3
	Dominance	5.063	5

It is curious how, in spite of considering a complex sound more dominant, players chose a simple one instead before playing the game. This decision was not uncommon, and was actually made by most participants in this experiment. Insights and opinions around the very concepts of valence, dominance or arousal when it comes to detecting spatial sound were varied, and every user ended up choosing what appealed to them most, intuitively. During the oral survey made after the experiment took place, when asked about their selection, 3 users mentioned "storms" or "thunder" as a reason for considering low tones helpful when trying to orient themselves. They thought those sounds were, in their own words, "easy to track", "full of energy" or "very deep". The rest of the participants had a similar reasoning behind their decisions, which were made with a strong base on personal experiences.

7.4 Data analysis and conclusions

One of the most notable pieces of information that can be extracted from the experiment described above is that the independent variable (the presence or absence of an attribute selector affecting music in the game) is statistically related to the difference in total time (t_3)

achieved by subjects during the experience.

Additionally, subjects in group A had a higher result in h_i, which means they perceived music as a helper more than participants in group B. It is important to note that precedents for this kind of behaviour have not been found in pertinent academic literature.

Another interesting observation that emerges from this experiment is that $t_n - t_{n-1}$ greatly increases with every measurement —that is, when every statuette was recovered. This is inversely proportional to the number of statuettes present in the map. It seems reasonable to think that the amount of time elapsed in finding a figure can increase when their remaining number is lower, because the probability of finding them by chance is also reduced. A need to backtrack and search more thoroughly also emerges when there are fewer objects to pick up. However, the increasing variation in average t_n between groups A and B (as is explained in section 7.3) points to another, more important correlation. If the fact that both prototypes (A and B) were identical except for the personalised music is taken into account, it should be possible to associate the differences in mean time and total time to the differences in audio.

Moreover, there existed some counterintuitive aspects in the results. For example: the lack of consistency between SAM test results and player preference when selecting attributes inside the prototype could be happening due to multiple reasons. This particular matter cannot be confidently answered with the amount of information that was retrieved during the described experiment, though several interesting possibilities arise as a result. This unexpected behaviour could be the result of a lack of correspondence between mental states when selecting sound attributes and when answering a test like the SAM. While said test is a more relaxed experience, which is not limited by time constraints, the video game asks players to concentrate much more, and gives them a clear goal that must be met in as few seconds as possible. As a result, it is possible that different attributes are found dominant in these different contexts, thus creating the mentioned variations in the results.

Another possible explanation for this data is that users learned to better identify dominant attributes through the duration of the SAM test, taking into account the specific variations in complexity, pitch and rhythm presented to them. This would mean the first answers would be less informed than the last ones, and that their decisions inside the final selector would imply a previous and meticulous "weighting up" of every possible option.

An inversion in the order of the test and attribute selection may shed light on how subjects choose depending on their predisposition. Besides, the SAM test only accounts for emotional scales (valence, arousal and dominance), while different measures could be needed to determine how easy to track a sound is for different persons, as sounds traditionally considered to be more dominant may not be necessarily easier to track for all users. Personal preference could actually be more important than dominance when it comes to finding sound sources in three-dimensional virtual environments.

Previous statements aside, one thing is clear: user capacity to select attributes and user

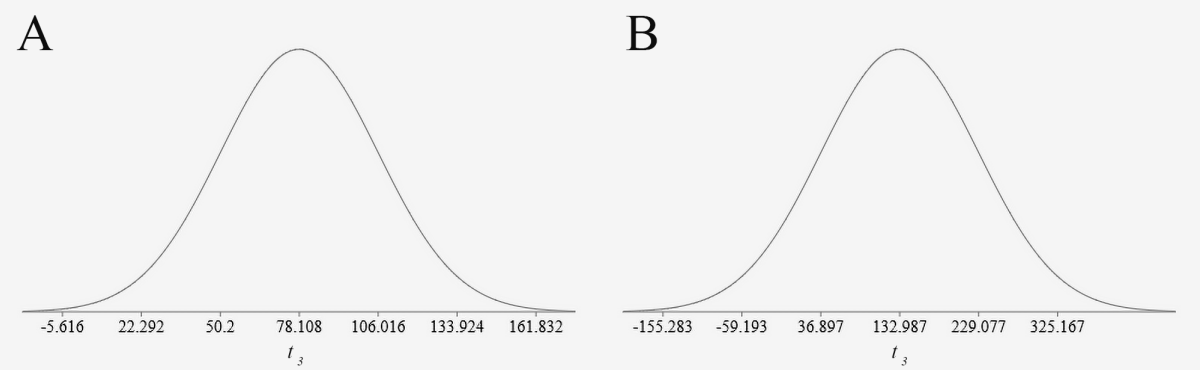


FIGURE 7.7. Bell curves for t_3 in groups A and B.

performance are, nevertheless, statistically related in this experiment. Consequently, it is possible to assert that the mere ability to choose correlates strongly with a lower average time of completion (t_3) in group A, when compared to group B in an experimental environment with the aforementioned conditions.

Some additional aspects of the retrieved data are worth mentioning, especially those concerning user distribution. Following the Central Limit Theorem [97], the presumption of normal distribution would only solidly apply to groups with a number of participants (N) equal or greater than 30. However, as can be appreciated in Figure 7.7, t_3 histograms form bell-like curves in both groups (A and B), even with less data, and the confidence interval of the mean is high enough (above 95 %) to trust the results. The lack of women in the sample is, however, a bigger issue. Only 4 out of 33 participants were females, which produces a genre bias and makes these conclusions only strictly applicable to men.

In summary, the most relevant conclusion that can be extracted from the attached data is that there exists an influence, which derives from the mere act of selecting preferred attributes, over player performance when solving a 3D labyrinth in the conditions established above.

This effect, whilst somewhat predictable, was not verified in the past in any other research, and opens the path for further exploring the consequences this relationship has in user behaviour in similar contexts.

The increase in performance achieved when using adaptive, spatialised music may also make a preference selection system like the one proposed here useful in virtual, interactive experiences where the inclusion of a GUI is not an option. The lack of a GUI is not uncommon in VR experiences, as was mentioned in Chapter 4, and sound could fill the void it leaves, at least in terms of behavioural game design.

Additionally, the variety in attributes selected by subjects in group A was also surprisingly high, and it points to the existence of a very complex population when it comes to auditory

preference.

This experimental phase laid the foundations of the next iteration in the development of LitSens: the implementation of an intelligent system integrated in the sound engine, as a way to adapt to player musical preferences and emotions without a previous test, which improved immersion while reducing even more the amount of GUI elements needed in a video game. This new development involved including a new logic for the automatic selection process in LitSens, while leaving the adaptive music engine untouched.

CHAPTER 8

ADAPTIVE MUSIC THROUGH GESTURAL INPUT IN LITSENS

"Beauty of scene; stateliness of movement; sweetness of sound — these are the graces that seem to reward the mind that seeks enjoyment purely for its own sake."

Virginia Woolf

In LitSens, adaptive music is conceived as a method for being consequent with player emotions, thus increasing presence, and this kind of adjustment to situations that arise during *gameplay* has to rely on information retrieved in real time to work properly.

In previous iterations of the LitSens architecture, said information was strictly textual and explicit. This limited the scope of interest of the system to situations where there was some form of dialogue happening. Therefore, a hypothetical implementation in a commercial video game would require a different approach, as the soundtrack ought to be consistent during a full gaming session, whether the different situations the player faces involve textual input or not.

In room-scale VR games, players wear an HMD (usually along a couple of controllers) which can provide lots of information about their position and posture, as well as their speed and acceleration rate when moving. The following sections explain how this information was

used to detect a particular emotion in players in real time, which allowed for adapting the soundtrack live.

8.1 A method for gestural analysis using machine learning

The amount of data retrieved by common HMD sensors is huge, and can become difficult to analyse for a human being in search of a pattern. This is why some machine learning techniques make sense in this scenario. Specifically, it was decided to use a multi-layer perceptron (MLP) because of its usefulness in classification prediction, when inputs can have labels assigned to them [14, 57, 60, 112]; afterwards, one-dimensional convolutional neural networks (1D CNN) were also utilised, because they allow for better predictions when variables are time-dependent [7, 124, 128].

8.1.1 Parameters

In this particular case, the parameters chosen for analysis were a series of variables which are easy to read and store when using a commercial HMD like the HTC Vive, and that give plenty of information about a player's posture in real time. The data matrix used for training and classification was composed by the following variables:

- Mouse axis X: Tracks the existence of movement in the X axis of the mouse. Returns a positive or negative value depending on the direction.
- Mouse axis Y: Tracks the existence of movement in the Y axis of the mouse. Returns a positive or negative value depending on the direction.
- Camera velocity X: Tracks the speed at which the camera moves in the X axis. Returns a negative or positive value in units per second.
- Camera velocity Y: Tracks the speed at which the camera moves in the Y axis. Returns a negative or positive value in units per second.
- Camera velocity Z: Tracks the speed at which the camera moves in the Z axis. Returns a negative or positive value in units per second.
- Camera rotation X: Tracks the absolute rotation value of the main camera in the X axis. Returns a value in degrees.
- Camera rotation Y: Tracks the absolute rotation value of the main camera in the Y axis. Returns a value in degrees.

– Movement direction X: Tracks the direction of the forward vector of the movement, and returns its X value.

– Movement direction Y: Tracks the direction of the forward vector of the movement, and returns its Y value.

– Movement direction Z: Tracks the direction of the forward vector of the movement, and returns its Z value.

The engine chosen for the making of this experience was the already-mentioned Unity, due to its high adaptability and modularity, and its potential for integrating a wide variety of machine learning implementations.

Only one emotion was used for the final training of the system: horror, as it is usually associated with strong and clear gestures and postures. Thus, elements in the aforementioned matrix were labelled as "scary" (with a value of 1) or "null" (with a value of 0).

Besides, the inclusion of variables for tracking both mouse and camera movement was decided so as to allow the system to work in first-person experiences without a HMD, as well as in VR. This was particularly useful during testing and debugging.

8.1.2 Implementing a neural network (MLP) through Scikit-learn in Unity

Due to its efficiency and the ease of use of the Python language, Scikit-learn's MLP[1] was chosen as the first neural network model that would be implemented in this project.

This learning algorithm has to be trained with a set of features $X = x_1, x_2, ..., x_n$ and a target y, and produces a function $f(\cdot) : R^n \longrightarrow R^o$, where n is the number of features and o is the number of dimensions for output. The activation function utilised in each layer in this case was a rectified linear unit (ReLU) function: $f(z) = max(0, z)$, so that $f(z)$ equals zero when z is less than zero. The reason for this was simply that, after testing a variety of activation functions, ReLu gave better results overall.

This Python library was connected to Unity by running on a different thread and being fed by the engine all the necessary parameters. A more detailed description of this system's architecture can be found in Figures 8.1 and 8.2.

As for the neural network itself, Figure 8.3 shows its structure. The input layer is formed by the 10 controlled variables, and produces 1 output with two possible values: "0" if a scare is not detected and "1" if it is. There are 3 hidden layers with 5, 4 and 3 neurons respectively. This configuration was reached after initial testing of the system, and is the one that gave better results in terms of emotion identification and training efficiency.

[1] `https://scikit-learn.org/stable/modules/neural_networks_supervised.html`

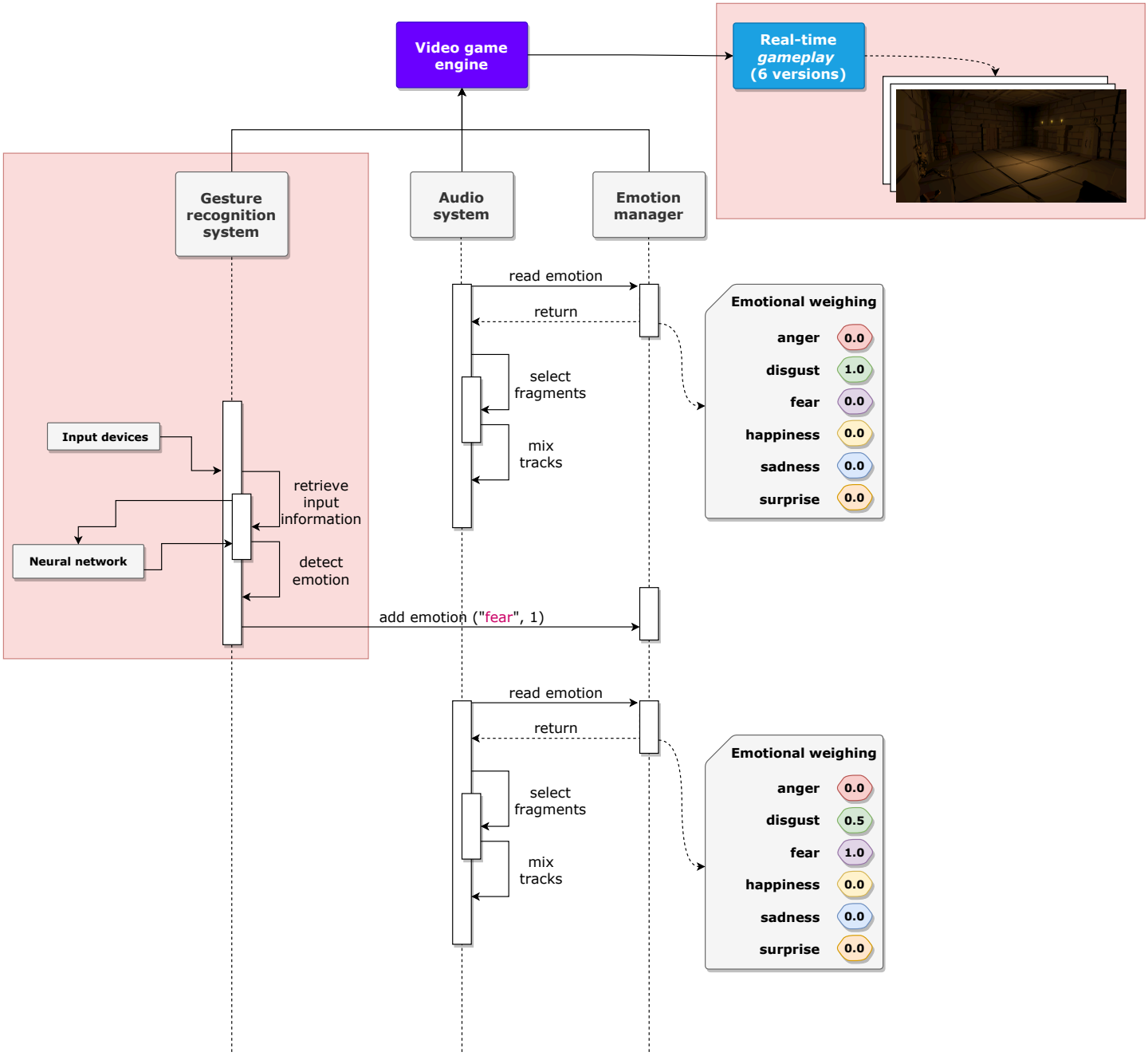


FIGURE 8.1. Data flow diagram for LitSens v0.4. Input information is now retrieved from a mouse and a keyboard or an HMD, and processed through a neural network. Emotional weights disappear in favour of binary outputs ("1" or "0"), given by the neural network. The "1" values represent the existence of the emotion the system has been trained to recognise. All changes from previous versions of LitSens are presented over a red background.

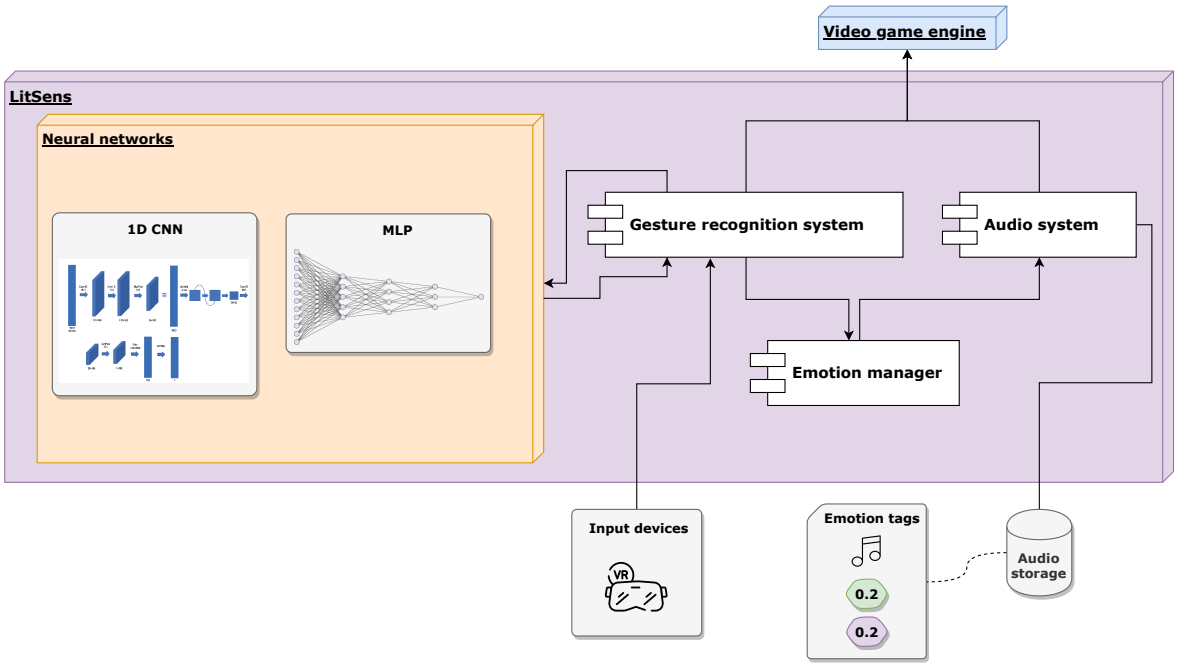


FIGURE 8.2. System diagram of LitSens v0.4. A new system is added to retrieve data from input devices and process it through neural networks.

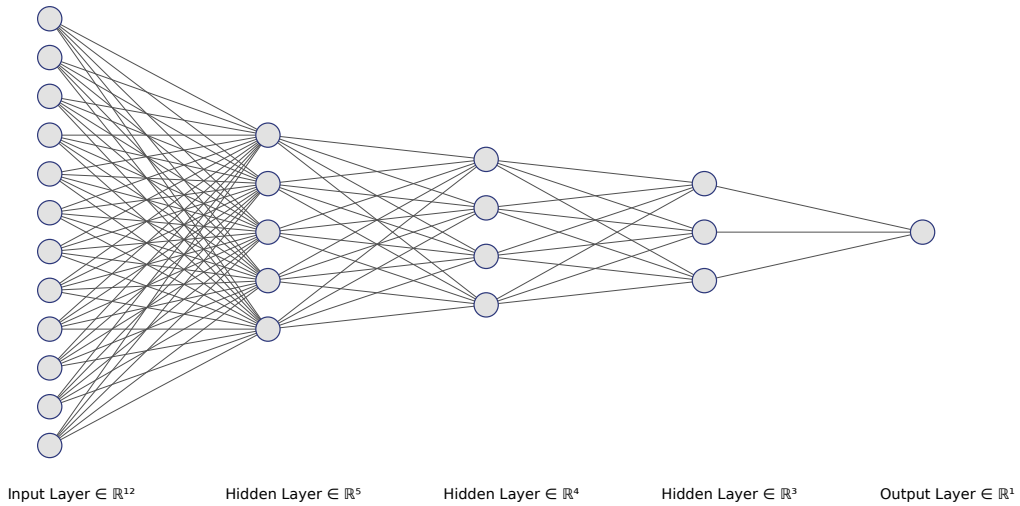


FIGURE 8.3. Structure of the multi-layer perceptron (MLP) utilised during the first phase of experimentation.

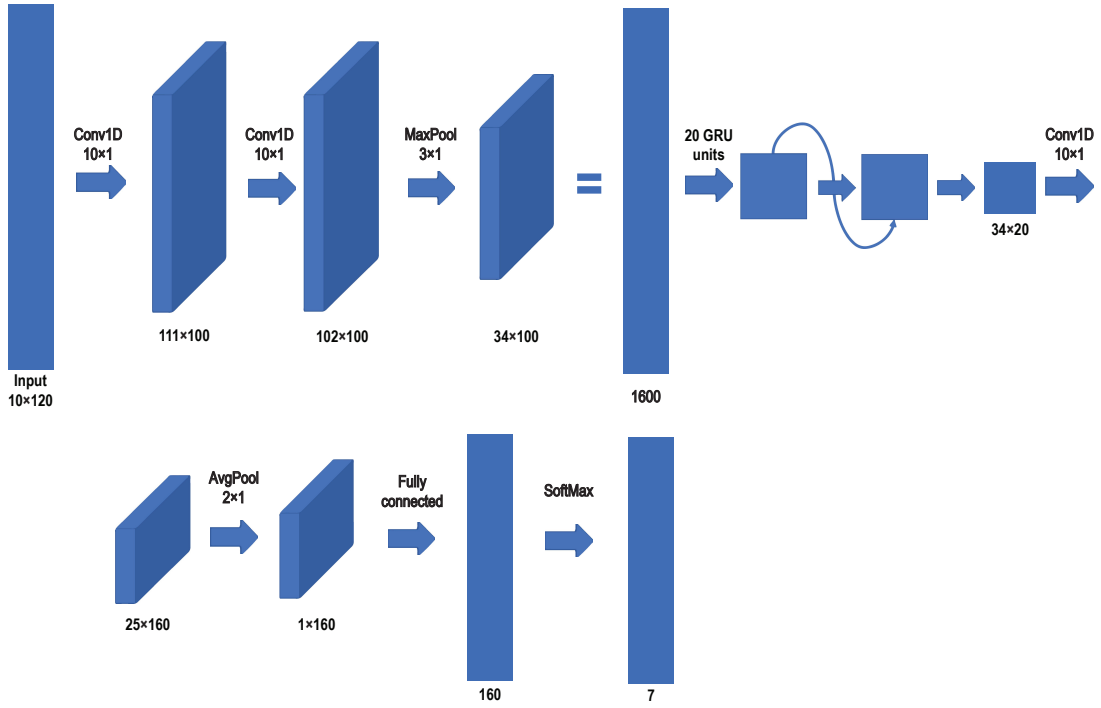


FIGURE 8.4. Structure of the one-dimensional convolutional neural network (1D CNN).

However, due to the nature of this initial implementation, the time window for emotion identification was too brief, and depended strictly on sudden gestures. To solve this problem, a different approach was taken, which involved passing the neural network a series of batches with samples from several seconds of *gameplay*, as is explained in the next subsection.

8.1.3 Implementing a 1D CNN through Keras and TensorFlow

Some human gestures and attitudes can be made evident by movements spanning more than a few milliseconds. The previously described implementation, based on a simple MLP, did not take measures over time, and instead relied on "snapshots" taken every 10 ms. In order to increase generalisation capacities of the neural network model present in LitSens, and reduce the number of random or false positives, periods of 12 seconds worth of data samples were fed to a one-dimensional convolutional neural network. This time window was decided upon after testing the system's performance with data extracted from real *gameplay* sessions.

This new implementation, depicted in Figure 8.4, sacrifices adaptiveness for consistency, as now audio can only change and adapt to player actions every 12 seconds. However, the

good results achieved during training make the system suitable for creating relevant audio atmospheres, which adapt to how a player behaves in a more general way during a gaming session.

Keras [2], an additional Python deep learning library, was utilised during this second phase, along with its TensorFlow[3] core.

Initially, 7 labels were included: one for each of the 6 basic emotions plus a "null" one. However, after realising the system was better suited for the identification of "horror", only 2 labels were utilised ("null" and "horror").

Kernel regularisation was added to every one-dimensional convolution layer represented in Figure 8.4, in order to avoid overfitting. During compilation, the optimiser was a simple stochastic gradient descent (SGD) with a learning rate of 0.001 and Nesterov momentum.

As Figure 8.4 shows, input data has a shape of 120 vectors of 10 elements each initially, but is then transformed into a 111 × 100 matrix due to the first convolution layer. Thus, the next input has a shape of 111 × 100 and, after a second convolution layer, changes into a 102 × 100 matrix, which will then be transformed by a pooling layer in order to reduce the amount of parameters, dividing them by 3. The resulting data (34 × 100) is transformed into a single vector with 1600 elements, enters a 20 unit Gated Recurrent Unit (GRU) layer and changes its dimensions to 34 × 20. The next convolution layer transforms the data into a 25 × 160 matrix. Next, an average global pooling layer transforms the 25 inputs into only one. Lastly, there is a fully connected layer with 160 elements. The output is processed with a softmax regression function, and produces a layer with 7 elements, which correspond to the 6 + 1 emotions described in the next section.

8.1.4 Prototyping

Initially, 6 prototypes were made in order to depict all of Paul Ekman's emotions. The idea behind each game scenario was to induce a single, very clear emotion in players, in order to train the AI. Figure 8.5 shows screenshots for each one of the environments, which shared a series of characteristics:

- The player spawns at the center of the virtual environment, and every relevant in-game element is visible from that position, as long as the camera is rotated.
- All environments are designed with room-scale VR in mind. When controlled with mouse and keyboard and no HMD, the player rotates the camera using a mouse, and walks using the "W", "A", "S" and "D" keys; when wearing a HMD, the player just walks and rotates at will inside the environment.

[2] https://keras.io/
[3] https://www.tensorflow.org/

– Movement is limited by a series of collisions that match the edge of the environment itself. Once reached, the player cannot walk in that direction anymore. While in VR, these limits are determined by the HTC Vive's Chaperone itself. The Chaperone is a software-aid that helps players avoid physical objects in the real world by drawing a grid that encloses the area where moving freely is "safe".

– While observing the environment, the player receives a series of visual and auditory stimuli that try to induce each relevant emotion.

– The experience in each environment lasts for a few minutes (depending on how each user plays), and then ends.

The differentiating elements included in each environment were the following:

FIGURE 8.5. From (a) to (f): Screenshots of the prototypes used to represent each of the 6 basic emotions.

– **(a) Happiness:** Key elements were bright colours, dancing characters, flowers, fireworks and major-mode music. The environment consisted of a small archipelago in the middle of the sea, full of grass and flowers.

– **(b) Horror:** Key elements were darkness or dim lighting, skulls and bones, artefacts of torture, small or oppressive environments and dark, chaotic music. The environment was a small room filled with broken cutlery, skeletons, an iron maiden, candles, a dusty fireplace and a small, closed door.

- **(c) Sadness:** Key elements were rain, sunset, clouds, darkness, death, stillness and slow, sad music. The environment was a modified version of the archipelago, and included a rain effect, several dark clouds, red lighting, a grey sky and two corpses floating in the water.
- **(d) Anger:** Key elements were noise, bright lights, chromatic aberration, red tint, high-pitched sound and pressing music. The already-mentioned archipelago was modified so that it included only untextured meshes, and a slow-motion effect was applied to the main camera. Besides, a high-pitched tone was included during the whole experience, and a fake system error popped up after a few seconds of *gameplay*.
- **(e) Surprise:** Key elements were lack of expectation, sudden stimuli and rhythmic, pressing music. The same environment that can be found in the "Happiness" prototype was utilised, with two exceptions: there were no dancing characters, and lightning fell after a few seconds of *gameplay*. This lightning was accompanied by a strong thunder sound.
- **(f) Disgust:** Key elements were slime, meat, insects, gurgling sounds and atonal music. The islands were substituted in this case by floating viscera, and a sound of buzzing flies was also included.

8.2 Training

In order to recover data and create a training and a test set for the 1D CNN, 6 persons were asked to play each of the prototypes described above, while information from camera movements was being recorded into a log file. One additional participant was placed in a neutral place (the empty room described in Chapter 5) so as to retrieve measurements from an environment that is not specifically designed to induce any particular emotion, and with no ambient music. These persons were only given the following instructions before commencing:

- You will appear in a small, virtual room when the game starts.
- It will be possible for you to move around so as to examine the different elements in that room.
- Once you think you have seen everything there is to see inside the room, tell your supervisor, who will stop the game for you.

After each participant finished, their log was recovered and stored with the others, and then assigned the relevant tag with the emotion.

From the 7 subjects who participated in this process, 3 were women and 4 were men, with ages ranging from 25 to 53 years old. They had no previous knowledge of the nature of the LitSens system, and all of them were born in Spain.

Both implementations of the neural networks were trained using the parameters described above, but the MLP used data from only one long gameplay session during which several scares happened. Each measure was taken every $\frac{1}{10}s$, thus producing a matrix with 100 elements every second. To test the system, initial training was carried out with data from seven *gameplay* sessions, which produced log files with 14868 values in the first phase, and 92070 values in the second phase. These sessions included several scares or uncomfortable situations that were pre-designed and tagged in advance, so as to be able to associate them with player movement and posture.

Figures 8.6 and 8.7 show slices of samples from the data in one of the variables used for training, separated by the tag assigned to each one: "fear" for those samples taken while that emotion was present in the experience, and "null" for samples that were not associated to any particular emotion.

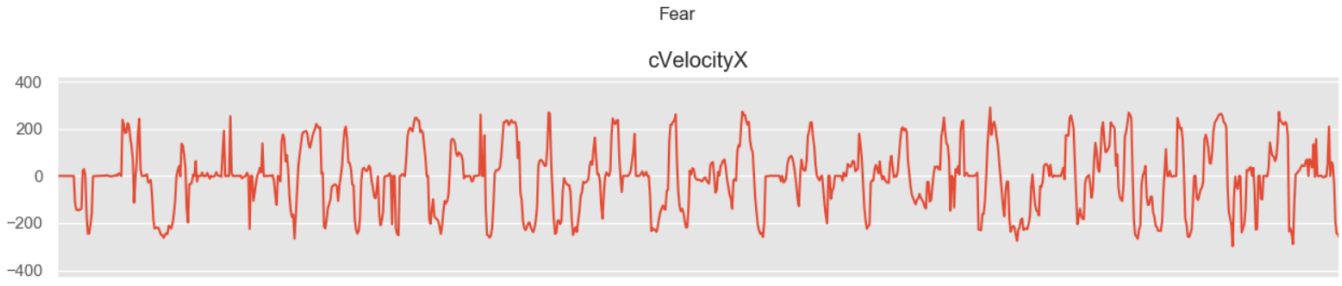


FIGURE 8.6. Slice of the training data for the variable "Camera velocity X", when an emotion tagged as "fear" was present.

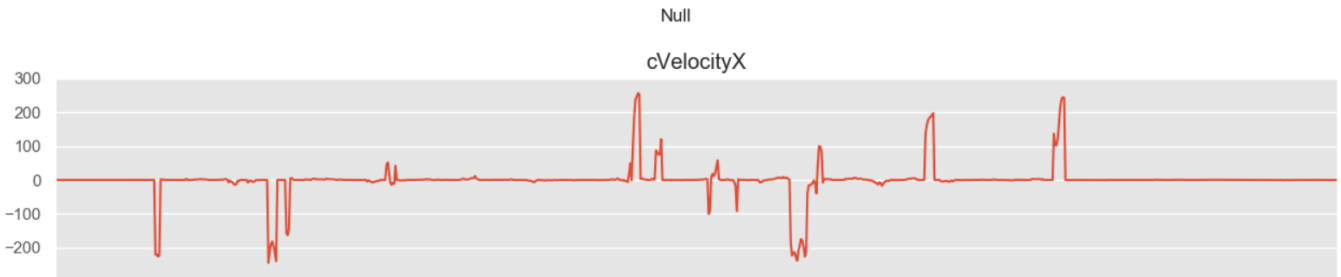


FIGURE 8.7. Slice of the training data for the variable "Camera velocity X", when no particular emotion was present.

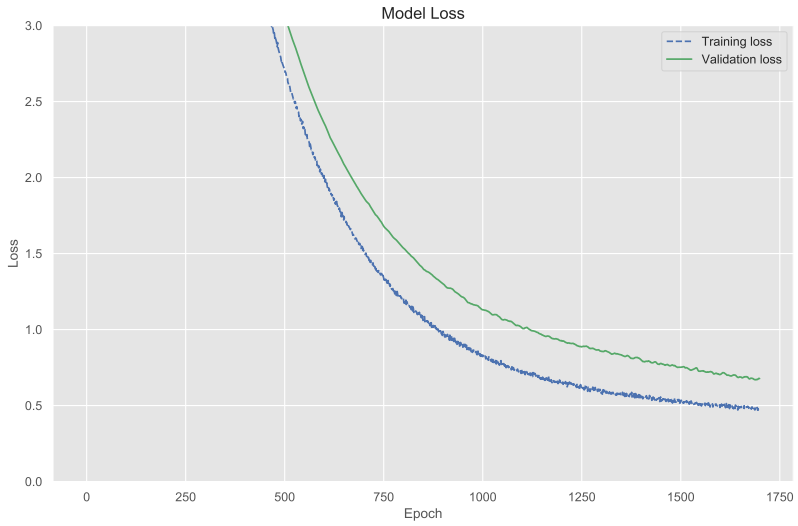


FIGURE 8.8. Evolution of the model loss for the 1D CNN implementation based on Keras and Tensorflow.

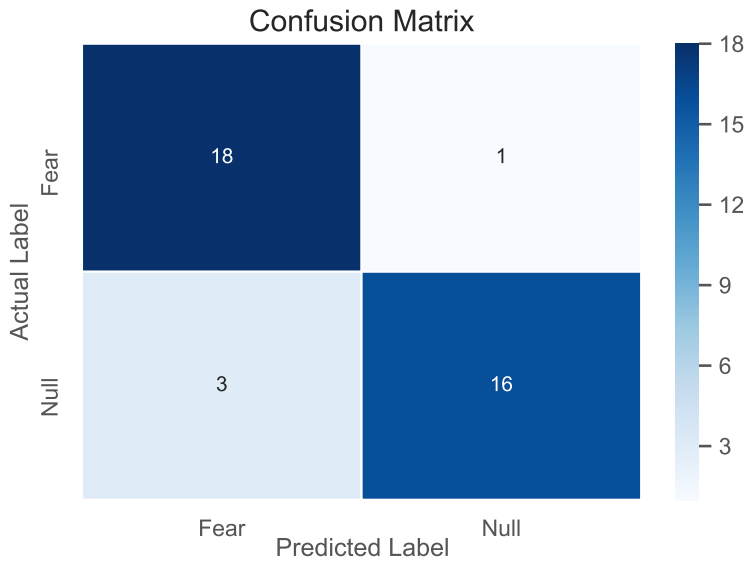


FIGURE 8.9. Confusion matrix for the test set from the 1D CNN implementation based on Keras and Tensorflow. A total of 45600 samples (with overlapping) was passed, in 38 batches of 12 seconds each. 19 batches were extracted from the set with no emotion tag, and the other 19 were from the set tagged with the emotion "fear".

During the first phase, the already described MLP was in charge of processing all data fed to it from the game engine through a plain text file.

In the second phase, when a 1D CNN was used, the system was trained with all six basic emotions; however, "fear" achieved the best results: 1700 epochs sufficed to reach the minimum loss (see Figure 8.8), and predicted labels from the test set were accurate (94.74 % success rate for "fear" and 84.21 % for "null"), as can be appreciated in Figure 8.9. A possible explanation for this situation is that the "fear" emotion is usually linked to more aggressive and apparent head gestures than the rest, like those which occur when suddenly scared or when being constantly alert. After observing the training sessions with real users, another factor was made evident: whereas emotions like happiness had very different and almost contradictory gestural representations by the same subject, fear was much more uniform, and consisted mostly of slight and sudden changes in head height, as well as quick head rotations in order to better control what was happening in the virtual environment. Besides, "fear" has revealed itself as the *less neutral* emotion, and extensively differentiates itself from the "null" category.

8.3 Hypotheses

The initial hypothesis behind this experiment stated that the system using a 1D CNN would correctly detect "fear" and "null" emotions through head or camera movements in a real first-person VR experience which did not share any data with the training or test sets. This would be checked through both the logs generated by the application and a human observer watching every session.

A second hypothesis presumed that, once an adaptive soundtrack was generated as a result of correctly detecting the relevant emotion, the sense of presence in subjects would increase when measured by a normalised version of the SUS test. This hypothesis would not reinforce the idea that the system is able to detect emotions and produce relevant changes in the soundtrack, but it would give an interesting insight on how much presence varies as a result of the described strategies.

8.4 Experiment design

After having the system trained, and during the posterior testing phase, 22 subjects were selected and distributed in groups of 11 members each (A and B), and they were assigned a slightly different version of the same experiment, which asked players to explore the room

that was tagged with the emotion "fear" during the previous phase. The only difference between A and B experiences was that the first one (A) was running LitSens along the 1D CNN and adapting audio to scares detected during gameplay, while the second one (B) only had non-adaptive ambient music. The MLP implementation was not tested during this phase, as the 1D CNN was considered an evolution of the system and a better option overall for detecting emotions over time. The possibility to use both methods at the same time in the future remains open, in spite of this, as they do not interfere with each other in any way and their performance is good.

Before beginning with the experiment, participants were given the following set of instructions on paper (translated from the Spanish version):

- You are going to play inside a virtual reality environment.
- Before commencing with the experiment, you must make sure that the headphones provided to you and the head-mounted display are correctly placed over your head.
- You will be able to move and walk freely inside the limits determined by a line that will appear on the floor when you approach a wall or an obstacle of any kind.
- It is not necessary to physically interact with anything inside the virtual environment. You will only need to visit it.
- You must look for an arrow in order to complete the experience.
- Once you have observed the arrow inside the virtual environment, tell your supervisor and remove the headphones and the headset.
- When the experiment is complete, you will have to fill in two forms, one in Spanish and one in English.
- Please, do not talk with other people about what you have seen or heard inside the virtual room after finishing.

During the experiment, the HMD used was an HTC Vive, along with a pair of over-ear headphones. The environment, in both its A and B versions, tried to scare users by flashing a red light and playing a screaming sound when they approached the arrow they had to find. There was no time limit for the mentioned task; users were allowed to roam freely for as long as they wanted.

Immediately after finishing, each subject was directed to a computer where they would take the SUS [106] presence test through a Python application which automated the process and normalised the total scores. When this test was completed, another form had to be filled in; this form was similar to the one described in the second section of Appendix D, and had

the goal of retrieving sociological information about participants. The only differences were that this survey did not ask about player performance in video games, musical ear or the usefulness of sound while playing; instead, it asked three questions about this particular experience:

– Have you been scared at some point of the experience?

  — Yes — No —

– Do you think the soundtrack has significantly changed or evolved during *gameplay*?

  — Yes — No —

– If you answered "yes" to the previous question, do you think the soundtrack was adapting to how scared you were?

  — Yes — No —

These questions were posed in order to make sure users were correctly perceiving how this new addition to LitSens influenced their experience.

8.4.1 Demography

During the experiment, there were a total of 22 participants, none of which had ever heard of LitSens or participated in the previous phase. There were 8 women and 14 men, distributed in two groups so that each had 4 women and 7 men. The only prerequisites these subjects had to meet to be able to take the experiment were to not have hearing problems and to speak English fluently.

As the experiment took place in a university in Spain, most participants (18) were born in this country, whereas 4 of them came from abroad (2 from Colombia, 1 from Chile and 1 from Mexico). The languages used during the experiment were Spanish (for the instructions and one of the surveys) and English (for the SUS presence test).

Average ages in group A and B were very similar: 28.91 and 29.27 respectively. The mode was 27 in group A and 28 in group B.

Only 3 participants did not have a university degree, as most subjects had obtained a Ph. D (3), a master's degree (12) or a regular degree (4).

When they were asked if they play games frequently, the results were very similar in both groups: an average of 3.09 out of 5 in a Likert scale was obtained in group A, and of 3 out of 5 in group B. The value 5 was meant for people who played games every day, and 1 represented subjects who did not play video games at all.

8.5 Results

Results from the tests that were passed after training and testing the system aim to validate how this last iteration of LitSens works, by establishing a comparison between groups A and B.

Firstly, there are clear differences between groups in the answers given to the three questions mentioned above, posed at the end of the experiment. When asked if they were scared at some point, 7 subjects from group A answered "yes", a response which was given by 4 members of group B. To the question "Do you think the soundtrack has significantly changed or evolved during *gameplay*?", all participants in group A (11) answered "yes", whereas only 2 gave the same answer in group B. Besides, 9 subjects in group A believed the soundtrack was adapting to how scared they were, while only 1 had the same opinion in group B. These questions were conceived as a way to validate that the experiment was working as intended, and the results can be considered positive, as most subjects noticed a difference in audio, and associated that difference to how they behaved in the game.

However, a second goal of this experiment was to check how the levels of presence varied between the two groups. As Table 8.1 and Figure 8.10 show, there exists a statistically significant difference in presence values between groups A (with an average value of 0.6623) and B (achieving a mean of 0.5065), determined by a Student's T-test which produces a p-value of 0.0379. This means presence is significantly higher in group A than in group B and, as the only difference between them was the gesture analysis system which provided adaptive sound, this difference can be statistically related to presence fluctuations.

TABLE 8.1. t Student's T-test for levels of presence achieved by users in groups A and B.

Group	**A**	**B**
Subjects (N)	11	11
Average	0.6623	0.5065
Std. Deviation	0.1539	0.1743
SEM	0.0464	0.0526
P-value	0.0379	

Logs collected during each session show scares were detected in 100 % of the first occurrences for every subject —all of them associated with a very explicit jump scare included in the experience. When a complete cycle of 12 seconds passed after this first scare, however, detection rates lowered in group B and maintained a high level in group A. From this information, it can be deducted that changes in group A's soundtrack caused this difference, reinforcing the feeling of fear in users and creating alertness.

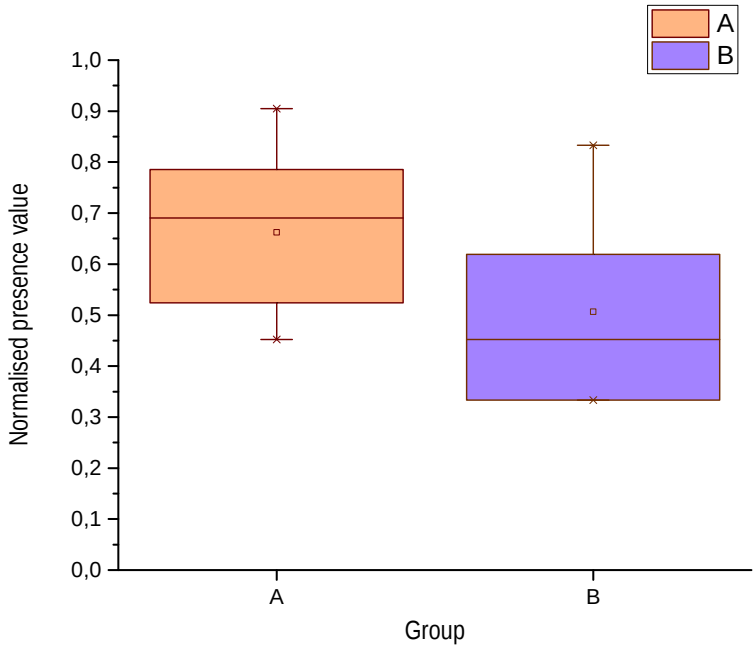


FIGURE 8.10. Differences in presence in both experimental groups (A and B).

8.6 Conclusions

The main conclusions that derive from this last experiment with LitSens are the following:

- People in group A felt more scared than people in group B during the experiment. Though more data would be needed to establish a firm relationship between this fact and the changes in music, this result is consistent with the rest of the data retrieved, as well as with previous research presented in this thesis [65, 70].

- All participants in group A were able to identify the changes in music in response to the emotion at hand. Additionally, they established a connection between the emotion and those changes, which means the system's perceived responsiveness is high.

- Lastly, the values of presence achieved by subjects in group A were notoriously higher than in group B when using the SUS test. This opens the door for further experimentation, as it points towards a statistical relationship between the addition of adaptive music and the presence experienced.

Additionally, an interesting behaviour was detected: auditory changes between the two versions of the experiment seemed to not have any particular influence over the first scare

suffered by subjects, but the next occurrences were reinforced if adaptive music was present. This also influenced the final answer given by participants to the question "Have you been scared at some point of the experience?". The superior number of positive replies in group A can be linked to the presence of adaptive music in it; in the case of most group B participants, a simple jump scare at the beginning of their experience was not enough to feel generally "scared".

CHAPTER 9

DISCUSSION AND CONCLUSIONS

"I have a story to tell you. It has many beginnings, and perhaps one ending. Perhaps not. Beginnings and endings are contingent things anyway."

Iain M. Banks

This thesis has swirled around many different approaches to multimodal interface design, with a focus on auditory interactions. In this process, several techniques (described through Chapters 3 to 8) were explored, and collaborated on the task of developing LitSens as an adaptive music system which takes into account player emotions and level of presence, while helping in-game design tasks like player orientation. This last chapter serves the purpose of analysing and discussing the main contributions of this research process, and of listing the conclusions that were reached in the end.

9.1 Discussion

Before delving into the conclusions themselves, it is important to take some time to reflect on the implications of the results presented through this work, as well as on their validity in larger, more complex populations or different virtual environments.

Chapter 3 introduced the subject of movement in VR while measuring simulator sickness and presence. The results, supported by a high correspondence between the SUS and TPI tests, anticipated a tendency of the video game industry of allowing locomotive motion in VR applications. However, these results cannot be extrapolated to a general population. Most subjects participating in these experiments were young men, and played first-person video games frequently, which means they were acclimatised to virtual movement in its many shapes. Further experimentation would be needed to determine if the effects of *walking* or *flying* in VR are similar in different populations. Still, all data retrieved is useful if the context of this research is taken into account, as this whole thesis focuses on improving the experience for users of, specifically, first-person and VR interactive applications.

Besides, the results from both SUS and TPI questionnaires could have been complemented by physiological measurements, such as an electroencephalogram (EEG), an electrocardiogram (ECG) or the measure of heart rate. These options were considered, but discarded due to lack of resources. Additionally, SS was not measured over time, due to the fact that a test could be properly passed only after each experience was finished. A continuous measurement of SS through physiological indicators would be appropriate to better analyse the fluctuations of this value over time, thus being able to associate them to certain in-game interactions.

This chapter, though unrelated to auditory interfaces, set the foundations for subsequent advancements by considering locomotive motion in VR a viable possibility. Without this research, as well as the support the game industry has been giving to these kind of virtual movement solutions in recent years, the development of an adaptive music system based on expressions of emotions through movements would not have been possible. Had the results been negative during this experimentation phase, a completely different approach would have been taken, possibly based on stationary VR experiences, or more traditional, non-VR first-person video games.

In Chapter 4, behavioural guidance by means of musical tracks was tested. Due to the neutrality of the virtual environment in which the experiment took place, it could be argued that the extracted guidelines would only work in this type of scenario. However, subsequent research based on these results [70] has produced similar reactions in players, even though the 3D virtual environments used during testing were bigger and more complex. In spite of everything, as also happened in the previous chapter, most participants in this experiment were frequent video game players, which makes these results applicable only to a narrow population.

The Bartle test itself has also received criticism due to its shallowness [8, 125] and the

excessive complexity of the "explorer" profile, which makes it wider than the rest of the categories. Nevertheless, during the experiment described in this thesis, the Bartle test was put to work with the only intention of creating four sets of users which shared emotional traits, from a very general perspective. The relatively even distribution between groups is also an indicator that the "explorer" group was not excessively unbalanced in this particular case, even though it was the most common option. Aside from that, the goal of this part of the experiment was to know if there existed a relationship between player profiles and behaviour when listening to certain types of music; even if there existed a bias in terms of player group size, the results would be valid, as long as there was a statistically significant amount of users in each set.

Chapter 5 presented a new sound spatialisation technique based on the application of LPFs to sounds coming from behind the player in a 3D environment. Though the results obtained were promising, a larger set of subjects would have benefited their consistency; the test having a high duration and being in-person drastically limited the amount of participants. Another interesting approach for this research process would have been to implement the new, LPF-based spatialisation system in a real video game. This was not done due to potentially very high costs; besides, there is a lack of popular open source video games which depend heavily on sound spatialisation.

The already-mentioned sex bias in the information technologies field also produced results that are mostly applicable to men. Women were distributed evenly between groups, but their small numbers could have had a decisive influence over the results. Moreover, as was mentioned in Chapter 5, gender differences in hearing have been discovered in previous research by Don, Ponton, Eggermont and Masuda [19], which means it could be an important factor in player performance in this scenario.

The conclusions reached during this experimental phase were successfully applied when designing auditory environments for the commercial video game *Song of Horror* (see Appendix E). This allowed for testing these methods in a real life scenario; they were considered more efficient, in computational terms, than traditional approaches to 3D audio, and reduced costs in terms of external software licenses due to being able to count on a proprietary solution.

In Chapter 7, LitSens was utilised to research new methods for guiding players in virtual, 3D environments, with a focus on auditory preference and sound dominance. After this research, a particular participant behaviour remains unexplained with currently available data: the lack of correspondence between the SAM test results and player preference when choosing certain sound attributes as more "dominant". A possible cause for this result, as was expressed in Chapter 7, is the lack of correspondence between the stimuli received by subjects in the two different environments where they were asked to choose dominant sounds, thus producing a change in mental states. Even though the task was similar, the environments where the experience takes place were very different: a traditional survey is much more

relaxed than a computer game where players are specifically asked to perform as efficiently as they can. Even though it has no particular interest for this research (as the focus is elsewhere), in the future it would be interesting to test how formal differences in the way a simple choice like this is proposed have influence over the final outcome.

The existence of an uncontrolled learning process cannot be discarded either, as the survey was passed before users had the opportunity to play the game. As a consequence, there is a possibility that the most informed decision was taken after subjects were "trained" by the first one.

Additionally, genre and age biases were also present in the experimental phase of this research process, due to the same reasons stated in the paragraphs above.

Lastly, in Chapter 8, a gestural recogniser based on neural networks was implemented in LitSens, with the intention of adapting music in real time to basic emotions expressed through head movements. This new development was based on a reasonable amount of data used for training; however, this data came from only 7 users, and a bigger sample would be needed for better generalisation. While testing this implementation, a similar problem arose: a total of 22 subjects participated in the experiment, and a bigger sample would be very useful to increase data confidence. Nonetheless, all data obtained is consistent with previous research articles, and points towards the same direction, which makes it more reliable, and useful for validating how LitSens works and its effects on people.

It is important to note that, in this last experiment, users from group A had very similar behaviours while playing the game. Most of them walked around for a while without feeling particularly alarmed or fearful, but then they triggered a *jump scare* specifically placed near a skeleton lying on the floor. A few seconds after this moment, music started to be modified as a consequence of how users moved their heads after the scare, and the attitude of subjects themselves changed until the experience was finished: they showed more signs of being alert, and triggered numerous positives for the emotion "fear" when their movements were processed through the neural network.

This means adaptive music in LitSens, in this case, does not act as an emotion trigger by itself, but rather supports an existing emotion and intensifies it. This fact makes the techniques used by LitSens useful when trying to enrich auditory atmospheres, but also when designing strong emotional sequences for video games. VR horror games could benefit from this technique extensively, because the intensity of fear can be adequately regulated through the soundtrack depending on how scared a user actually is.

An interesting way to extend this research would be to work on the responsiveness of the neural network, in terms of time. In its current state, LitSens can adapt correctly to emotions being detected during gameplay, but with a 12 second delay. The inclusion of a different set of parameters could alter the density of the retrieved data, and thus, how quickly the 1D CNN can identify emotions.

9.2 Conclusions

This Ph. D. thesis was conceived with the intention of improving presence and immersion in first-person and VR virtual worlds, enriching the overall experience. It started with an open, multidisciplinary approach, and slowly aimed towards sound design by virtue of the good results achieved in this field.

Though a variety of software was built for testing and experimenting, LitSens stands as the culmination of all the described research efforts: it works as an adaptive music system which takes into account player emotions, and makes use of neural networks to read these emotions in user gestures.

Since this research started, a series of conclusive findings have been reached:

- Free *walking* or *flying* movement is possible in VR without simulator sickness being a determining factor in decreasing presence, as is explained in Chapter 3.

- The field of game design can benefit from including semantic value in auditory interfaces. Free roaming in VR cannot rely completely on GUIs due to their negative impact on presence, while player orientation through audio is not intrusive and works reasonably well (see Chapters 4, 5 and 7).

- There is a lack of conscious attention to sound in virtual applications, but audio can have a big impact on user behaviour, notwithstanding actual awareness, as seen in Chapter 4.

- Changing the following musical parameters has an influence over emotional perception of a soundtrack in subjects: tone, rhythm, similarity between overlapping track waveforms, channel panning, general harmony, general frequency range and structural complexity (frequent variations in rhythmic and tonal patterns). A more in-depth explanation of every item in this list is available in Chapter 4.

- The impact sound has over *gameplay* is statistically related to psychological profiles and player emotions, as seen in Chapter 4.

- The addition of LPFs to sounds coming from outside the FOV in 3D virtual worlds can improve recognition of audio source location in environments similar to the ones described in Chapter 5, along with player score, measured in seconds, when completing a simple positioning task.

- Adapting three basic sound parameters (tone, rhythm and complexity) to player preference increases their performance when recognising these sounds in a labyrinthine 3D environment (see Chapter 7).

– Linking head gestures and emotional responses is possible in LitSens (following the experiment detailed in Chapter 8), and can be used to reinforce emotional states by applying relevant auditory stimuli.

The main contributions of the present thesis to this field of research can then be summarised as:

– The building and testing of LitSens: an adaptive music system with a focus on emotions. Different versions of LitSens have been implemented in two of the most popular video game engines: Unity and Unreal Engine. Its multiplatform architecture allows this software to work on any desktop computer (Mac or PC), on popular game consoles (PS4, Xbox One or Switch) and on mobile devices (Android phones and tablets).

– Establishing a relationship between non-accelerated movement and a lack of simulator sickness in VR applications, which allows for free roaming in that kind of virtual environments without decreasing presence.

– Presenting successful methods for player orientation in 3D environments through auditory interfaces.

– Determining the relationship between head (camera) gestures and emotions in first-person video games, and building a neural network capable of detecting fear in players. This detection was utilised to adapt LitSens' audio tracks in a seamless way that does not require any explicit interaction with the player.

As a consequence of these contributions, several new and interesting research opportunities are open. LitSens has the potential to grow both as a tool for adaptive music in video games and as an instrument for research. The addition of a bigger database of human-made fragments can increase the performance of the system and make it apt for commercial video games of different genres, and it would be interesting to test how presence changes in users when switching musical genres or styles, or when applying this technology to longer, emotionally complex experiences. Training the gesture recogniser with bigger samples of users would also be appropriate, in order to increase its confidence and detection rates. Besides, new sensors can potentially be added to increase precision, such as the ones present in Oculus Touch or HTC Vive controllers, or in external devices (electrocardiograms, electrodermal activity sensors, electroencephalograms, etc.).

The present research has also had a tight relationship with the field of sound design for video games, understood as a branch of the game design discipline. Academic contributions found in this text happened in a context of state-of-the-art technology in constant evolution, and some of them were applied to commercial video game projects, such as *Song of Horror* (see Appendix E).

Ultimately, this thesis intends to bring new attention to the link that exists between human emotions and sound, and to how this link can help design rich and meaningful interactive experiences which do not focus only on representation accuracy, but also on how people interact with video games as a whole.

APPENDIX A

Appendix A: Soundtrack influence questionnaire

The following questionnaire is a translation of the one used during initial research for this thesis. It contains questions destined to distinguish among different sociological and psychological profiles, and also includes a test to determine the influence of two types of soundtrack on a simulated narrative.

1. State your age:
   — 6 to 10 years old
   — 11 to 14 years old
   — 15 to 18 years old
   — 19 to 25 years old
   — 26 to 35 years old
   — 35 to 45 years old
   — More than 45 years old
2. How often do you play video games?
   Never - 1 - 2 - 3 - 4 - 5 - Every day
3. How important do you think graphical quality is in a video game?
   Not important - 1 - 2 - 3 - 4 - 5 - Very important

4. How important do you think the soundtrack is in a video game?

   Not important - 1 - 2 - 3 - 4 - 5 - Very important

5. How important do you think the story is in a video game?

   Not important - 1 - 2 - 3 - 4 - 5 - Very important

6. How important do you think game design is in a video game?

   Not important - 1 - 2 - 3 - 4 - 5 - Very important

7. How important do you think audio is in a video game?

   Not important - 1 - 2 - 3 - 4 - 5 - Very important

8. In which of these video game genres do you think the story plays an important role?

   — Adventure

   — Puzzle

   — Action

   — Driving & sports

   — Platformer

   — Role Playing Game (RPG)

   — Horror

   — Strategy

   — Social simulator

9. In which of these video game genres do you think the soundtrack plays an important role?

   — Adventure

   — Puzzle

   — Action

   — Driving & sports

   — Platformer

   — Role Playing Game (RPG)

   — Horror

   — Strategy

   — Social simulator

10. In which of these video game genres do you think graphical quality plays an important role?
    — Adventure
    — Puzzle
    — Action
    — Driving & sports
    — Platformer
    — Role Playing Game (RPG)
    — Horror
    — Strategy
    — Social simulator
11. In which of these video game genres do you think game design plays an important role?
    — Adventure
    — Puzzle
    — Action
    — Driving & sports
    — Platformer
    — Role Playing Game (RPG)
    — Horror
    — Strategy
    — Social simulator
12. In which of these video game genres do you think audio plays an important role?
    — Adventure
    — Puzzle
    — Action
    — Driving & sports
    — Platformer
    — Role Playing Game (RPG)
    — Horror
    — Strategy
    — Social simulator

13. What is your favourite video game genre?

    — Adventure

    — Puzzle

    — Action

    — Driving & sports

    — Platformer

    — Role Playing Game (RPG)

    — Horror

    — Strategy

    — Social simulator

14. Which of these statements define you better?

    — I don't like to drop out of things. Whenever I start something, I finish it. Besides, I'm well organised.

    — I don't pursue success, but knowledge. I prefer to understand others thant to be able to influence them.

    — I am good at communicating what I feel to others. I like to help the people around me, and I enjoy social encounters.

    — I like to develop my abilities and show them to others. I am a competitive person, and I frequently compare myself with the rest of the people in my environment.

15. From the following virtues, choose the one that represents you better:

    — Intelligence.

    — Empathy.

    — Determination.

    — Pragmatism.

16. From the following flaws, choose the one that represents you better:

    — Fickleness.

    — Pride.

    — Selfishness.

    — Anger.

While you play a first-person video game, you find yourself in a dark room, facing two closed doors. The following soundtrack is playing*:

*The use of headphones is highly recommended to ensure an optimal listening experience. Before starting, ensure that they are correctly placed and that you are in a quiet environment.

A piece of music in stereo was included here. The left and right channels played variations of the same song with similar characteristics.

1. What do you think there is behind each door?

   (Introduce two words or groups of words, separated by a comma)

2. Which door would you open first?

3. Imagine there is a source of white light coming from under one of the doors, and a source of red light coming from under the other. Which one would be the white door?

   — The left one

   — The right one

Imagine the same scene described above while you play this other track:

A piece of music in stereo was included here. The left and right channels played variations of the same song: one of them was sadder and slightly more complex, whereas the other one was happier and simpler.

1. What do you think there is behind each door?

   (Introduce two words or groups of words, separated by a comma)

2. Which door would you open first?

3. Imagine there is a source of white light coming from under one of the doors, and a source of red light coming from under the other. Which one would be the red door?

   — The left one

   — The right one

4. If your answer to any of these questions has changed between the first and second iteration, briefly explain why.

APPENDIX B

Appendix B: TPI, SSQ and SUS questionnaires

The following versions of the TPI, SSQ and SUS tests were used to build applications which passed them automatically and applied weighting when necessary. These were run through the terminal, as can be seen in Figure B.1.

TPI Questionnaire

Thank you very much for agreeing to complete this questionnaire.

The questions on these pages ask about the media experience you just had. This may have been watching television, watching an IMAX or Omniverse film, or using a virtual reality (VR) system.

There are no right or wrong answers; please simply give your first impressions and answer all of the questions as accurately as possible, even questions that may seem unusual or to not apply to the particular media experience you just had. For example, in answering a question about how much it felt like you were "inside the environment you saw/heard," base your answer on your feeling rather than your knowledge that you were not actually inside that environment.

Throughout the questions, the phrases "the environment you saw/heard" and "objects, events, or people you saw/heard" refer to the things or people that were presented in the media experience, not your immediate physical surroundings (i.e., the actual room you were in during the media experience).

Please circle the responses that best represent your answers. All of your responses will be kept strictly confidential.

How much did it seem as if the objects and people you saw/heard had come to the place you were?

Not at all 1 2 3 4 5 6 7 Very much

How much did it seem as if you could reach out and touch the objects or people you saw/heard?

Not at all 1 2 3 4 5 6 7 Very much

How often when an object seemed to be headed toward you did you want to move to get out of its way?

Never 1 2 3 4 5 6 7 Always

To what extent did you experience a sense of 'being there' inside the environment you saw/heard?

Not at all 1 2 3 4 5 6 7 Very much

To what extent did it seem that sounds came from specific, different locations?

Not at all 1 2 3 4 5 6 7 Very much

How often did you want to or try to touch something you saw/heard?

Never 1 2 3 4 5 6 7 Always

Did the experience seem more like looking at the events/people on a movie screen or more like looking at the events/people through a window?

Like a movie screen 1 2 3 4 5 6 7 Like a window

How often did you have the sensation that people you saw/heard could also see/hear you?

Never 1 2 3 4 5 6 7 Always

To what extent did you feel you could interact with the person or people you saw/heard?

None 1 2 3 4 5 6 7 Very much

How much did it seem as if you and the people you saw/heard both left the places where you were and went to a new place?

Not at all 1 2 3 4 5 6 7 Very much

How much did it seem as if you and the people you saw/heard were together in the same place?

Not at all 1 2 3 4 5 6 7 Very much

How often did it feel as if someone you saw/heard in the environment was talking directly to you?

Never 1 2 3 4 5 6 7 Always

How often did you want to or did you make eye-contact with someone you saw/heard?

Never 1 2 3 4 5 6 7 Always

Seeing and hearing a person through a medium constitutes an interaction with him or her. How much control over the interaction with the person or people you saw/heard did you feel that you had?

None 1 2 3 4 5 6 7 Very much

During the media experience how well were you able to observe…:

…the body language of the people you saw/heard?

Not well 1 2 3 4 5 6 7 Very well

…the facial expressions of the people you saw/heard?

Not well 1 2 3 4 5 6 7 Very well

…changes in the tone of voice of the people you saw/heard?

Not well 1 2 3 4 5 6 7 Very well

…the style of dress of the people you saw/heard?

Not well 1 2 3 4 5 6 7 Very well

How often did you make a sound out loud (e.g., laugh, speak) in response to someone you saw/heard in the media environment?

Never 1 2 3 4 5 6 7 Always

How often did you smile in response to someone you saw/heard in the media environment?

Never 1 2 3 4 5 6 7 Always

How often did you want to or did you speak to a person you saw/heard in the media environment?

Never 1 2 3 4 5 6 7 Always

To what extent did you feel mentally immersed in the experience?

Not at all 1 2 3 4 5 6 7 Very much

How involving was the media experience?

Not at all 1 2 3 4 5 6 7 Very much

How completely were your senses engaged?

Not at all 1 2 3 4 5 6 7 Very much

To what extent did you experience a sensation of reality?

Not at all 1 2 3 4 5 6 7 Very much

How relaxing or exciting was the experience?

Very relaxing 1 2 3 4 5 6 7 Very exciting

How engaging was the story?

Not at all 1 2 3 4 5 6 7 Very much

For each of the pairs of words below, please circle the number that best describes your evaluation of the media experience.

Impersonal	1	2	3	4	5	6	7	Personal
Unsociable	1	2	3	4	5	6	7	Sociable
Insensitive	1	2	3	4	5	6	7	Sensitive
Dead	1	2	3	4	5	6	7	Lively
Unresponsive	1	2	3	4	5	6	7	Responsive
Unemotional	1	2	3	4	5	6	7	Emotional
Remote	1	2	3	4	5	6	7	Immediate

Please indicate how much you disagree or agree with each statement below.

	Strongly Disagree						Strongly Agree
The way in which the events I saw/heard occurred is a lot like the way they occur in the real world.	1	2	3	4	5	6	7
The events I saw/heard could occur in the real world.	1	2	3	4	5	6	7
It is likely that the events I saw/heard would occur in the real world.	1	2	3	4	5	6	7

Overall, how much did the things and people in the environment you saw/heard…:

…sound like they would if you had experienced them directly?
Not at all 1 2 3 4 5 6 7 Very much

…look like they would if you had experienced them directly?
Not at all 1 2 3 4 5 6 7 Very much

…smell like they would if you had experienced them directly?
Not at all 1 2 3 4 5 6 7 Very much

Overall, how much did touching the things and people in the environment you saw/heard feel like it would if you had experienced them directly?
Not at all 1 2 3 4 5 6 7 Very much

How much did the heat or coolness (the temperature) of the environment you saw/heard feel like it would if you had experienced it directly?
Not at all 1 2 3 4 5 6 7 Very much

These next questions are about the media experience as a whole.

Have you ever seen the media presentation/experience you had today before?
____ No ____ Yes

How personally relevant was the content of the media experience to you?
Not at all 1 2 3 4 5 6 7 Very much

How was the picture quality during the media experience?
Very poor 1 2 3 4 5 6 7 Very good

How was the sound quality during the media experience?
Very poor 1 2 3 4 5 6 7 Very good

How comfortable were you with your viewing position?
Not at all 1 2 3 4 5 6 7 Very much

Overall, how satisfying or enjoyable was the media experience you just had?
Not at all 1 2 3 4 5 6 7 Very much

Please use the space below to provide your comments about the media experience.

_____

_____

_____

You're almost done! These last questions are about you. Again, all of your responses will be kept strictly confidential, so please answer as accurately and honestly as possible.

How old are you (in years)? ____ ____

Please indicate your gender: ____ Male ____ Female

What is your race?

____ Asian
____ African American
____ Hispanic
____ Pacific Islander
____ White
____ Other: _____

What is your level of education?

____ Some high school
____ High school degree
____ Some college
____ College degree
____ Some graduate school
____ Graduate school degree

What is your occupation? _____

How many hours do you spend watching television (including watching videotapes and DVDs) in a typical day? (please estimate as closely as possible)

_____

What size television set do you most often watch?

____ Pocket TV (less than 6 inches measured diagonally)
____ Small TV (6 - 18 inches measured diagonally)
____ Medium TV (19 - 21 inches measured diagonally)
____ Large TV (22 - 27 inches measured diagonally)
____ Extra large TV (regular tube type)(28 - 35 inches measured diagonally)
____ Large screen projection TV (bigger than 35 inches measured diagonally)
____ NOT SURE

How often do you use a video game system (at home, work, school, or at an arcade)?

____ Never
____ Less than once a month
____ 1-4 times a month
____ 5-10 times a month
____ 11-20 times a month
____ More than 20 times a month

How many times have you used an interactive virtual reality system (e.g., Virtuosity)?

____ Never
____ 1 time
____ 2-4 times
____ 5-7 times
____ 8 or more times

How much do you know about broadcast or film production?

None 1 2 3 4 5 6 7 A lot

THANK YOU <u>VERY</u> MUCH FOR COMPLETING THIS QUESTIONNAIRE.
WE TRULY VALUE AND APPRECIATE YOUR TIME AND EFFORT !!
PLEASE RETURN THIS QUESTIONNAIRE TO THE STUDY COORDINATOR

No_____ Date_____

SIMULATOR SICKNESS QUESTIONNAIRE

Kennedy, Lane, Berbaum, & Lilienthal (1993)***

Instructions : Circle how much each symptom below is affecting you right now.

1. General discomfort	None	Slight	Moderate	Severe
2. Fatigue	None	Slight	Moderate	Severe
3. Headache	None	Slight	Moderate	Severe
4. Eye strain	None	Slight	Moderate	Severe
5. Difficulty focusing	None	Slight	Moderate	Severe
6. Salivation increasing	None	Slight	Moderate	Severe
7. Sweating	None	Slight	Moderate	Severe
8. Nausea	None	Slight	Moderate	Severe
9. Difficulty concentrating	None	Slight	Moderate	Severe
10. « Fullness of the Head »	None	Slight	Moderate	Severe
11. Blurred vision	None	Slight	Moderate	Severe
12. Dizziness with eyes open	None	Slight	Moderate	Severe
13. Dizziness with eyes closed	None	Slight	Moderate	Severe
14. *Vertigo	None	Slight	Moderate	Severe
15. **Stomach awareness	None	Slight	Moderate	Severe
16. Burping	None	Slight	Moderate	Severe

* Vertigo is experienced as loss of orientation with respect to vertical upright.

** Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea.

Last version : March 2013

***Original version : Kennedy, R.S., Lane, N.E., Berbaum, K.S., & Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology, 3*(3), 203-220.

Simulator Sickness Questionnaire***

Kennedy, Lane, Berbaum, & Lilienthal (1993)***

Validation of the French-Canadian version of the SSQ developed by the UQO Cyberpsychology Lab :

- Total : items 1 to 16 (scale of 0 to 3).
  - « *Nausea* » : items 1 + 6 + 7 + 8 + 12 + 13 + 14 + 15 + 16.
  - « *Oculo-motor* »: items 2 + 3 + 4 + 5 + 9 + 10 + 11.

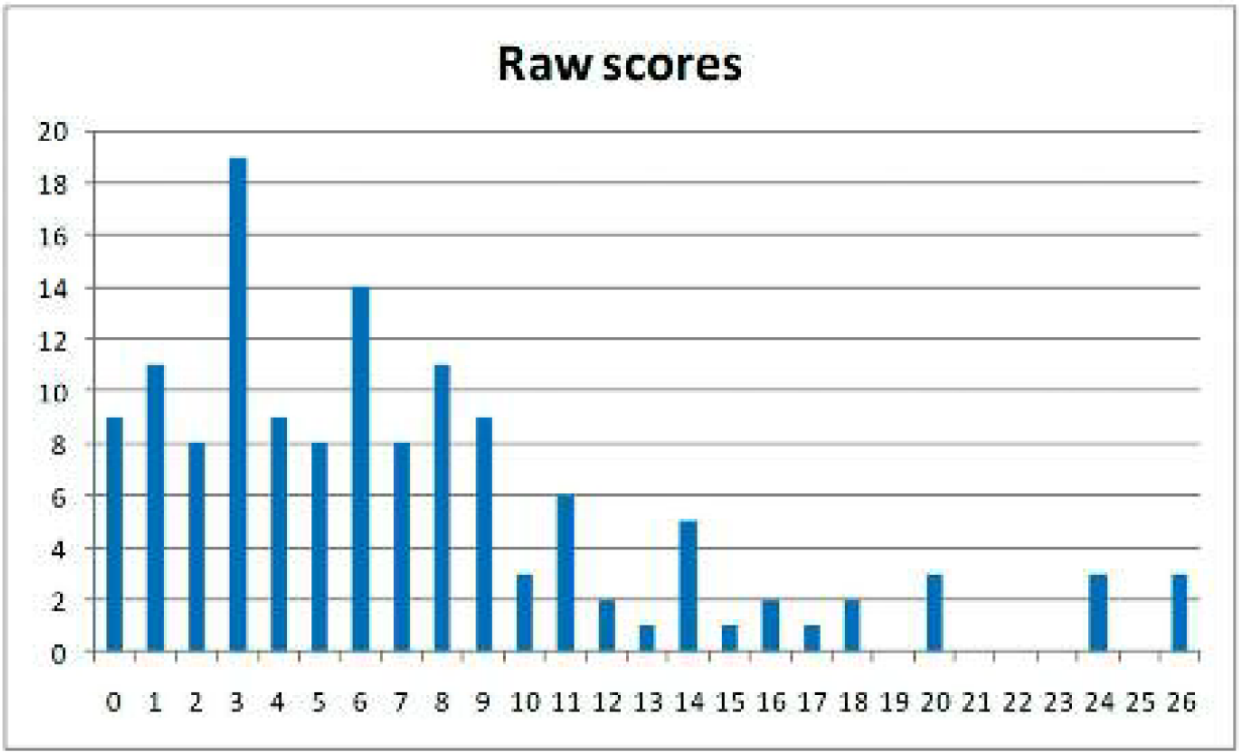
Please refer to the following articles for more information about the French-Canadian validated version :

BOUCHARD, S., Robillard, & Renaud, P. (2007). Revising the factor structure of the Simulator Sickness Questionnaire. Acte de colloque du *Annual Review of CyberTherapy and Telemedicine*, *5*, 117-122.

BOUCHARD, S., St-Jacques, J., Renaud, P., & Wiederhold, B.K. (2009). Side effects of immersions in virtual reality for people suffering from anxiety disorders. *Journal of Cybertherapy and Rehabilitation*, 2(2), 127-137.

BOUCHARD, S. Robillard, G., Renaud, P., & Bernier, F. (2011). Exploring new dimensions in the assessment of virtual reality induced side-effects. *Journal of Computer and Information Technology*, *1*(3), 20-32.

Based on results from Bouchard, St-Jacques, Renaud, & Wiederhold (2009), below are the mean scores reported:



Note. For the original scoring version, consult : Kennedy, R.S., Lane, N.E., Berbaum, K.S., & Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology*, *3*(3), 203-220.

SLATER-USOH-STEED QUESTIONNAIRE (SUS)

1. Please rate your *sense of being in the* virtual environment, on a scale of 1 to 7, where 7 represents your *normal experience of being in a place.*

2. To what extent were there times during the experience when the virtual environment was the reality for you?

3. When you think back to the experience, do you think of the virtual environment more as *images that you saw* or more as *somewhere that you visited*?

4. During the time of the experience, which was the strongest on the whole, your sense of being in the virtual environment or of being elsewhere?

5. Consider your memory of being in the virtual environment. How similar in terms of the *structure of the memory* is this to the structure of the memory of other *places* you have been today? By 'structure of the memory' consider things like the extent to which you have a visual memory of the virtual environment, whether that memory is in colour, the extent to which the memory seems vivid or realistic, its size, location in your imagination, the extent to which it is panoramic in your imagination, and other such *structural* elements.

6. During the time of your experience, did you often think to yourself that you were actually in the virtual environment?

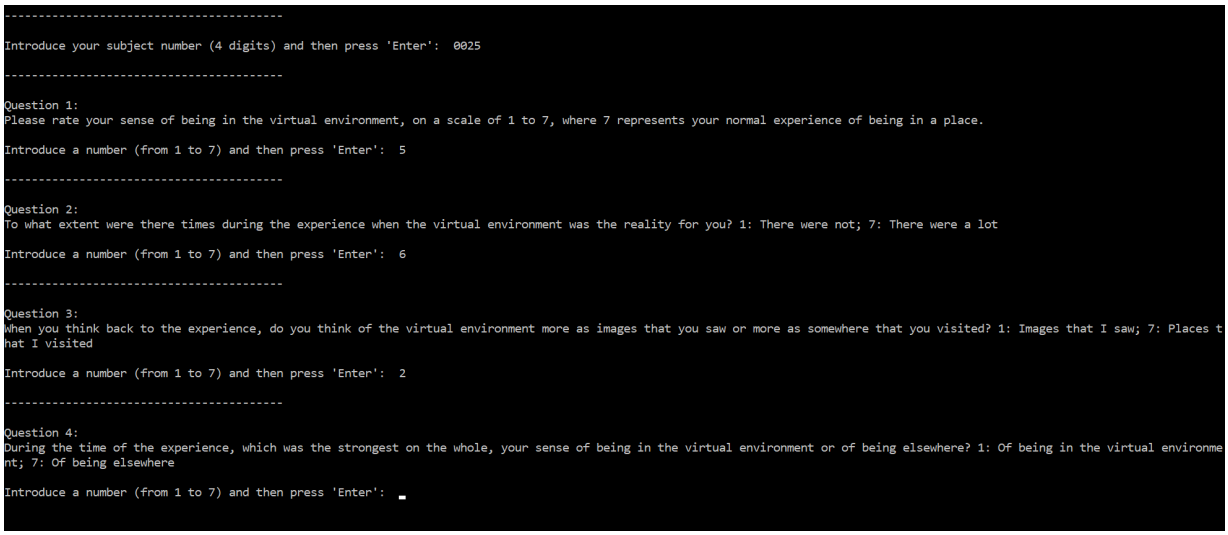


FIGURE B.1. Automatised SUS test running in Python.

APPENDIX C

APPENDIX C: SOUND SPATIALISATION TESTS

Tests used during the experiments described in Chapter 5 are attached here. The first one was passed after both versions of the first experiment, and had to be completed remotely. The second one was passed after an in-person session managed by a supervisor, and complemented the results obtained through two versions of the Unity prototype.

Sound spatialisation test 1:

Welcome to this survey for the research group NIL (Complutense University of Madrid). Please, read all questions carefully, and answer them with honesty.

During this experiment, it is essential that you use headphones; also, try to stay in a calm and quiet environment until you finish. Also make sure that your headphones are placed correctly over your head, so that the left speaker rests over your left ear, and the right one over your right ear.

All the information retrieved by this form will be treated confidentially, and used exclusively for academic purposes.

Thank you very much for your collaboration.

When you are ready to begin, press the "Next" button.

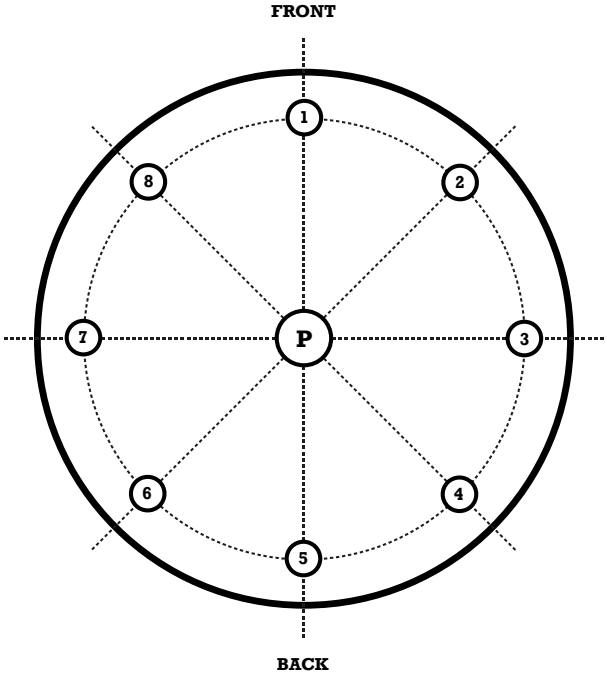
——

Through this experiment, you will have to identify the position of a given sound that will play towards you. The identifiable sound will always be the same one.

Here you can find an example of that sound. Once you have listened to it several times and are confident you will remember it, press the "Next" button.

An alarm sound could be played here.

The following diagram will be shown next to each question. It consists of a zenithal view of the player (P), surrounded by a series of locations (1 to 8). Think of the screen of your computer as the FRONT, and of your back as the REAR.



When you listen to the sound that was presented to you above, you will have to place it in the diagram, taking into account the position it comes from. For example: if you think the sound comes from the rear, you will have to place it in number 5, whereas if it comes from your right, you would have to select number 3, and so on.

When you are ready to begin, press the "Next" button.

Now listen to this audio fragment and indicate, by order of appearance, the positions from where you think the four key sounds come.

One of the audio tracks could be played here. An instance of the diagram was also included.

— Position of sound 1:
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - Unidentified

— Position of sound 2:
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - Unidentified

— Position of sound 3:
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - Unidentified

— Position of sound 4:
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - Unidentified

___

Now answer these brief questions as precisely as possible.

— How old are you?

_____

— State your gender:
- Woman
- Man
- Other: _____

— What is the highest level of education you have achieved?
- Elementary school
- High school
- Diploma
- College degree
- Master's degree
- Ph. D.

— Have you ever been diagnosed with auditory health diseases?
- Yes
- No

— Do you have a cold or feel congested?

- Yes

- No

Now indicate how much do you agree with the following statements, on a scale from 1 to 5.

— I often listen to music:

Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I have good hearing:

Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I have a "good ear" for music:

Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I often play video games:

Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— It is important for a game to have good audio:

Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

Sound spatialisation test 2:

Please, answer the following questions only after finishing with the experiment.

— State your subject number:

_____

— How old are you?

_____

— State your gender:
- Woman
- Man
- Other: _____

— What is the highest level of education you have achieved?
- Elementary school
- High school
- Diploma
- College degree
- Master's degree
- Ph. D.

— Have you ever been diagnosed with auditory health diseases?
- Yes
- No

— Do you have a cold or feel congested?
- Yes
- No

Now indicate how much do you agree with the following statements, on a scale from 1 to 5.

— I often listen to music:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I have good hearing:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I have a "good ear" for music:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I often play video games:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— It is important for a game to have good audio:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I found it easy to identify the position of sounds during this experiment:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I agree with the position that was shown to be right for each sound:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

APPENDIX D

APPENDIX D: SAM & ADDITIONAL TESTS

The following version of the Self-Assessment Manikin Test was passed to every participant in the experiment described in Chapter 7. An additional test passed alongside is also included, as well as both versions of the instructions for said experiment.

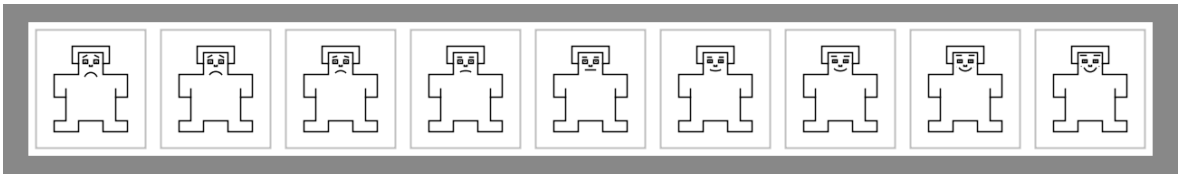
Self-Assessment Manikin Test:

— Please, state your participant number:

_____

Sound 1

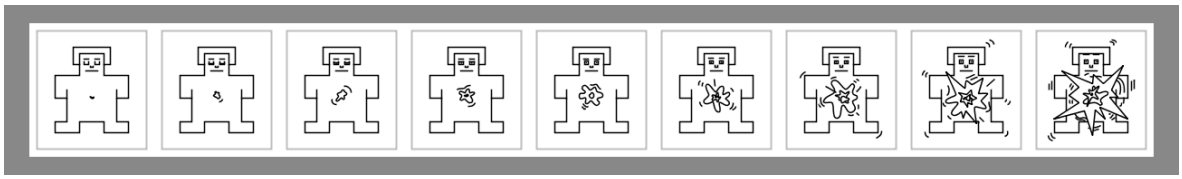
— Sound 1 valence:



— Sound 1 valence:

Negative - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Positive

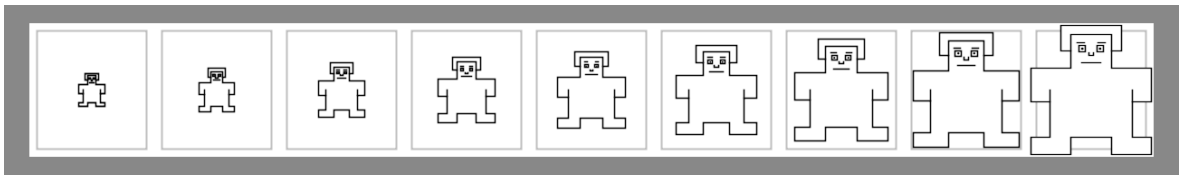
— Sound 1 arousal:



— Sound 1 arousal:

Calm - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Aroused

— Sound 1 dominance:

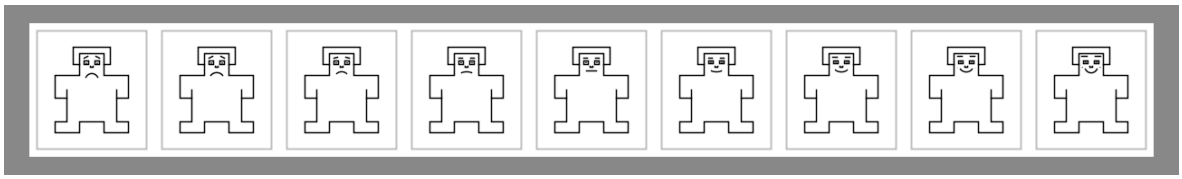


— Sound 1 dominance:

Insignificant - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Dominant

Sound 2

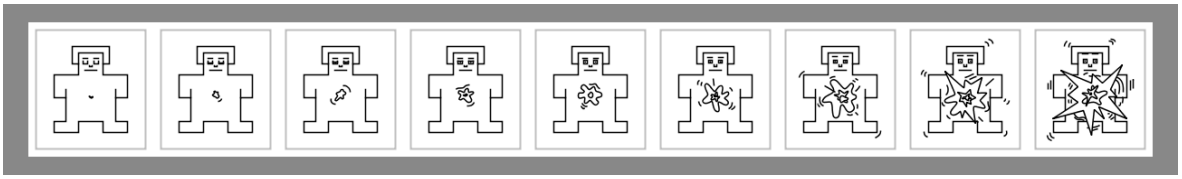
— Sound 2 valence:



— Sound 2 valence:

Negative - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Positive

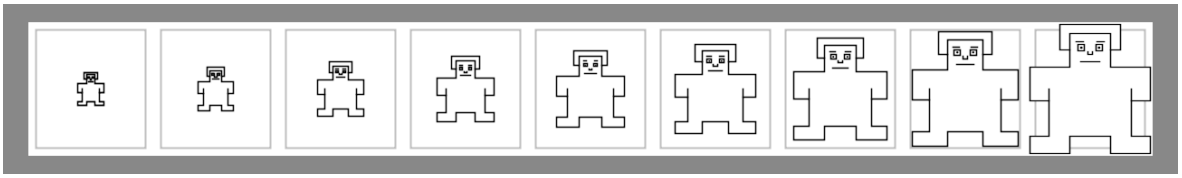
— Sound 2 arousal:



— Sound 2 arousal:

Calm - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Aroused

— Sound 2 dominance:

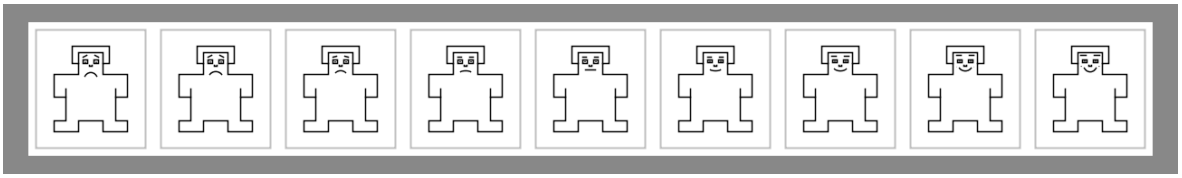


— Sound 2 dominance:

Insignificant - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Dominant

Sound 3

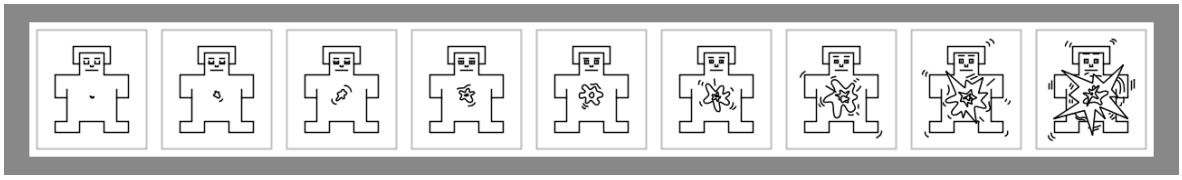
— Sound 3 valence:



— Sound 3 valence:

Negative - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Positive

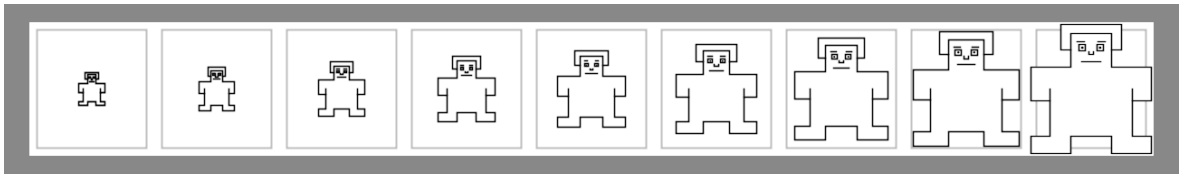
— Sound 3 arousal:



— Sound 3 arousal:

Calm - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Aroused

— Sound 3 dominance:

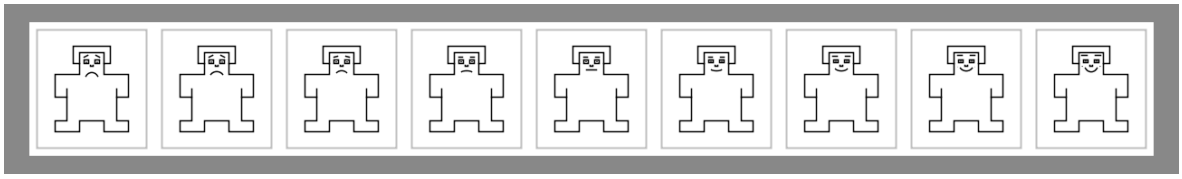


— Sound 3 dominance:

Insignificant - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Dominant

Sound 4

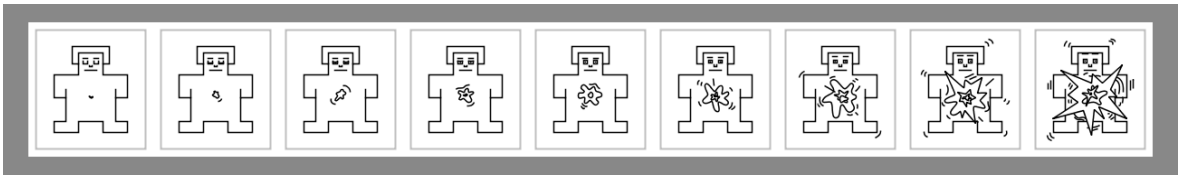
— Sound 4 valence:



— Sound 4 valence:

Negative - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Positive

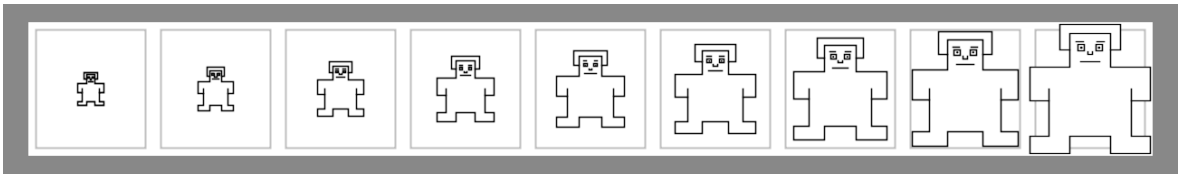
— Sound 4 arousal:



— Sound 4 arousal:

Calm - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Aroused

— Sound 4 dominance:

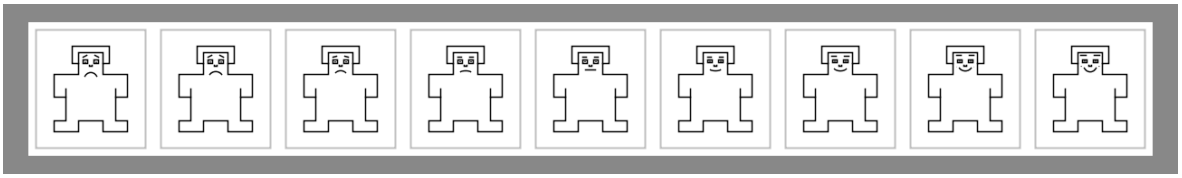


— Sound 4 dominance:

Insignificant - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Dominant

Sound 5

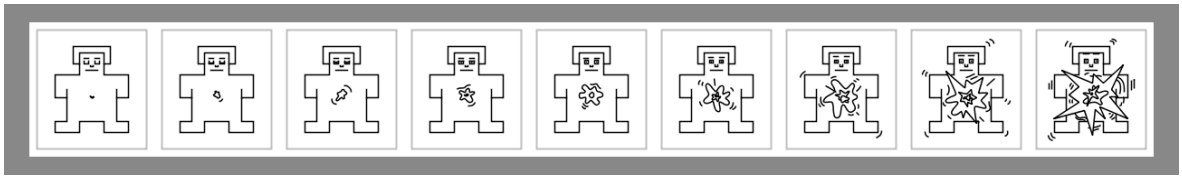
— Sound 5 valence:



— Sound 5 valence:

Negative - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Positive

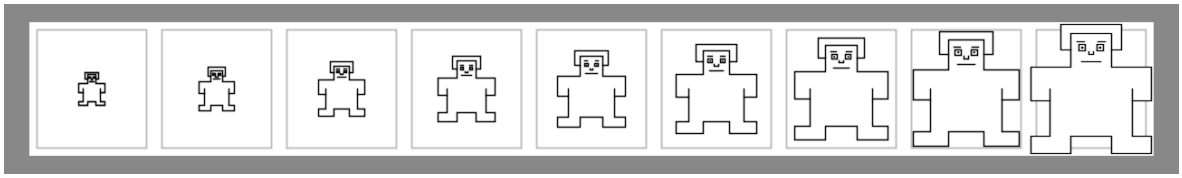
— Sound 5 arousal:



— Sound 5 arousal:

Calm - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Aroused

— Sound 5 dominance:

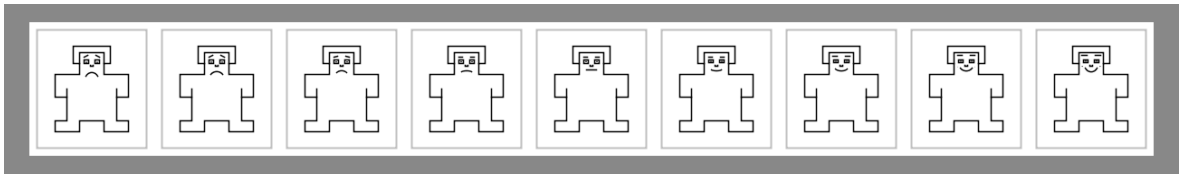


— Sound 5 dominance:

Insignificant - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Dominant

Sound 6

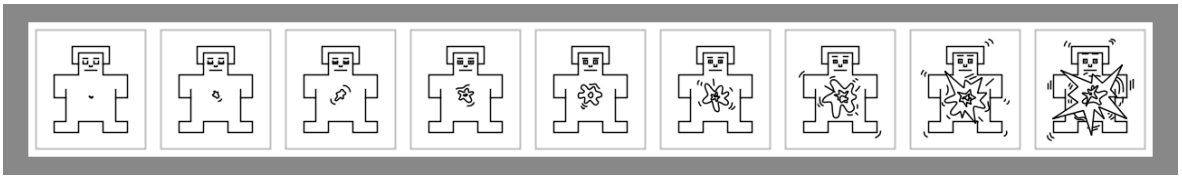
— Sound 6 valence:



— Sound 6 valence:

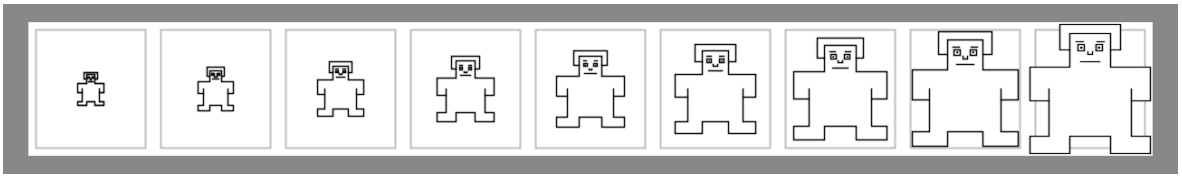
Negative - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Positive

— Sound 6 arousal:



— Sound 6 arousal:
Calm - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Aroused

— Sound 6 dominance:



— Sound 6 dominance:
Insignificant - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - Dominant

Sociological profile test:

Please, answer the following questions in an honest manner and press the "Send" button when you are done.

— State your participant number:

_____

— How old are you?

_____

— State your gender:
- Woman
- Man
- Other: _____

— What is your country of birth?

_____

— What is the highest level of education you have achieved?
- Elementary school
- High school
- Diploma
- College degree
- Master's degree
- Ph. D.

— Have you ever been diagnosed with auditory health diseases?
- Yes
- No

— Do you have a cold or feel congested?
- Yes
- No

Now indicate how much do you agree with the following statements, on a scale from 1 to 5.

— I often listen to music:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I have good hearing:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I have a "good ear" for music:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I often play video games:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I am good at playing video games:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— I am good at playing FP (first-person) games:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— It is important for a game to have good audio:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

— Audio from the game I have just played has helped me meet the goal of retrieving all figures:
Completely disagree - 1 - 2 - 3 - 4 - 5 - Completely agree

Experiment instructions (A version):

1. Please, sit down, take a deep breath and concentrate on this task. Wear these headphones and make sure they are correctly placed over your ears.

2. Before starting with the experiment, you will have to take a test during which you will have to indicate what type of emotions are awoken by three pairs of sounds. Even though you will listen to sounds in pairs, the emotional evaluation of each one will be separated from the rest.

3. Please, do not forget to include your participant number when asked to do so.

4. During the next test, you will have to evaluate sounds using three different scales: valence (sadness-happiness), arousal (calmness-arousal) and dominance (insignificance-dominance). With that purpose in mind, you will have to select, for each scale, the image that better depicts how you feel when listening to the sound at hand. If you

have any doubts about this process, please ask your supervisor before starting with the experiment.

5. After finishing this test, you will have to open LitSens' main application and select a combination of elements that you think make a sound stand out over the rest. It is important that you think about your options here thoroughly, and that you end up selecting the combination that, from your personal standpoint, is the most prominent or remarkable.

6. Once you press the "Play" button, you will have to control the movement of a character in a three-dimensional virtual environment. You will have a first-person perspective during the whole experience, and the only possible actions are moving on the floor in 4 directions (using the W, A, S and D keys) and rotating the camera (360 degrees) with the mouse. The layout of the virtual environment the player will have to walk through is similar to a labyrinth. The player's goal is to find and retrieve a total of 3 statuettes as quickly as possible. To pick up one statuette, you will have to place the player over them.

7. Once the experience has ended, and your completion time has been registered, you will have to answer another test, destined to know your sociological profile. It is important that you include your participant number in the required field, and that it is the same as the one you introduced during the first survey. It is also of utmost importance that you answer all questions.

8. Thank you very much for your collaboration!

Experiment instructions (B version):

1. Please, sit down, take a deep breath and concentrate on this task. Wear these headphones and make sure they are correctly placed over your ears.

2. Before starting with the experiment, you will have to take a test during which you will have to indicate what type of emotions are awoken by three pairs of sounds. Even though you will listen to sounds in pairs, the emotional evaluation of each one will be separated from the rest.

3. Please, do not forget to include your participant number when asked to do so.

4. During the next test, you will have to evaluate sounds using three different scales: valence (sadness-happiness), arousal (calmness-arousal) and dominance (insignificance-dominance). With that purpose in mind, you will have to select, for each scale, the image that better depicts how you feel when listening to the sound at hand. If you have any doubts about this process, please ask your supervisor before starting with the experiment.

5. After finishing this test, you will have to open LitSens' main application and let your supervisor select a series of parameters.

6. Once you press the "Play" button, you will have to control the movement of a character in a three-dimensional virtual environment. You will have a first-person perspective during the whole experience, and the only possible actions are moving on the floor in 4 directions (using the W, A, S and D keys) and rotating the camera (360 degrees) with the mouse. The layout of the virtual environment the player will have to walk through is similar to a labyrinth. The player's goal is to find and retrieve a total of 3 statuettes as quickly as possible. To pick up one statuette, you will have to place the player over them.

7. Once the experience has ended, and your completion time has been registered, you will have to answer another test, destined to know your sociological profile. It is important that you include your participant number in the required field, and that it is the same as the one you introduced during the first survey. It is also of utmost importance that you answer all questions.

8. Thank you very much for your collaboration!

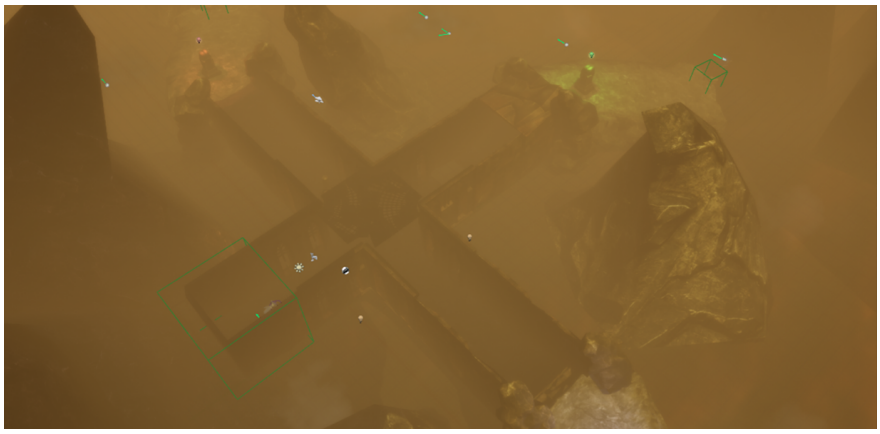
APPENDIX E

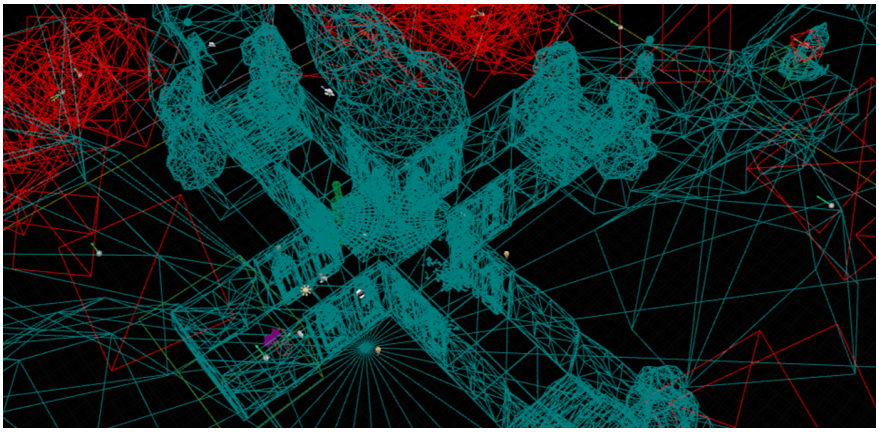
Appendix E: List of developed software

The following software applications were built as part of the present research process.

E.1 SS and presence prototype in Unreal Engine 4

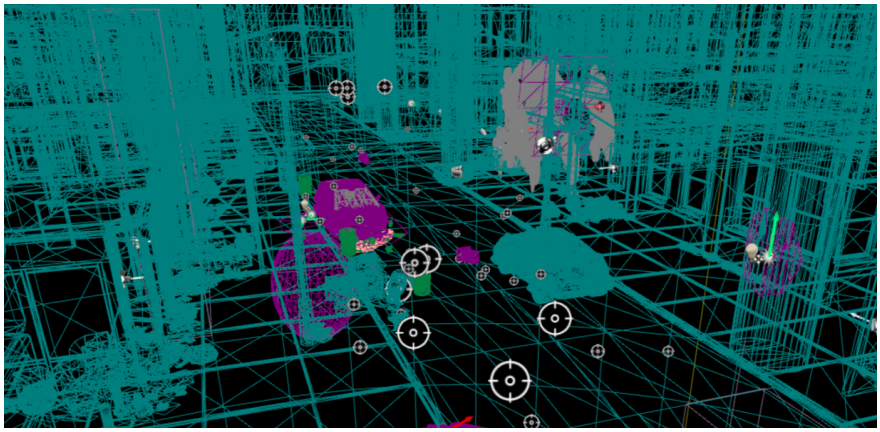
A first-person video game prototype was built exclusively for testing the relationship between presence, SS and movement. It included a small puzzle that had to be solved for the experience to end. More details can be found in Chapter 3.





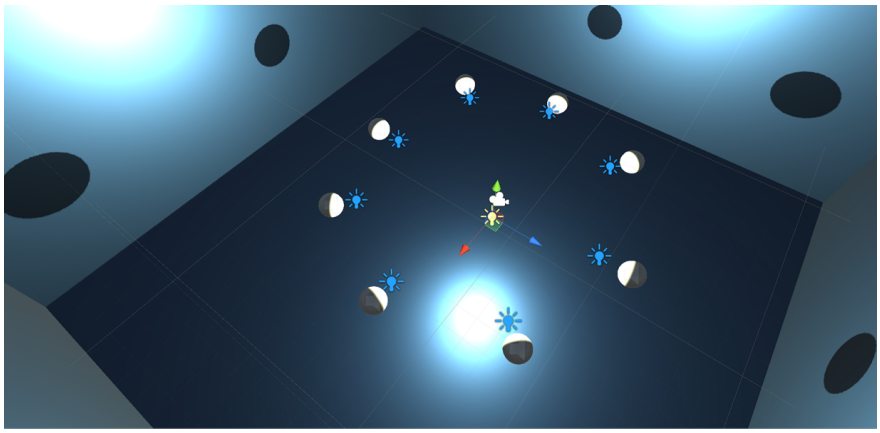
E.2 Sound positioning in *Showdown VR demo*

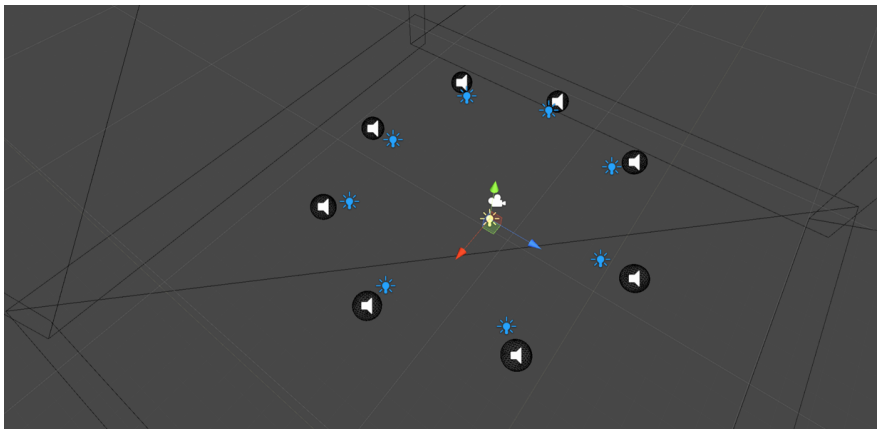
This free demo made available to developers by Epic Games was modified so as to include sounds coming from a total of 8 different positions around the player, and was used for preliminary testing before building the system described in the next section. A 3D pointer system was also added, so that test subjects could select different positions in the virtual space, depending on where they thought sounds were coming from at each time. This pointer was controlled through a HTC Vive controller, and the whole application was experienced in VR.



E.3 Sound positioning prototype in Unity

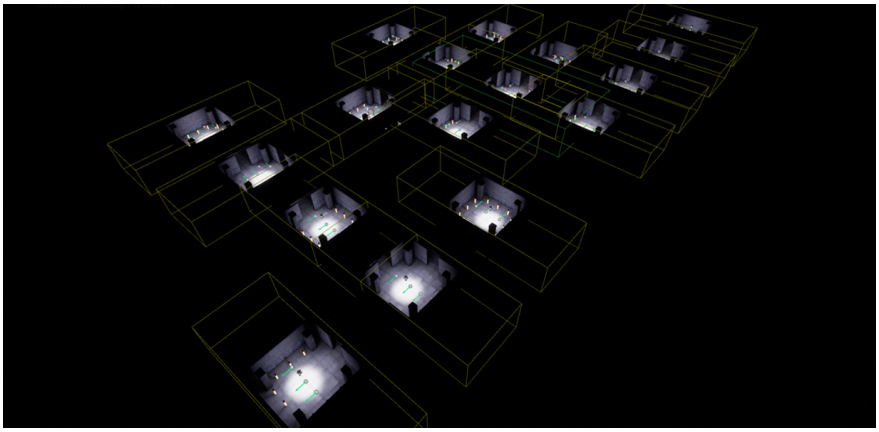
Initial experimentation with the *Showdown* demo was promising, and a standalone application was built with the purpose of experimenting in a more controlled environment, as can be seen in Chapter 5.

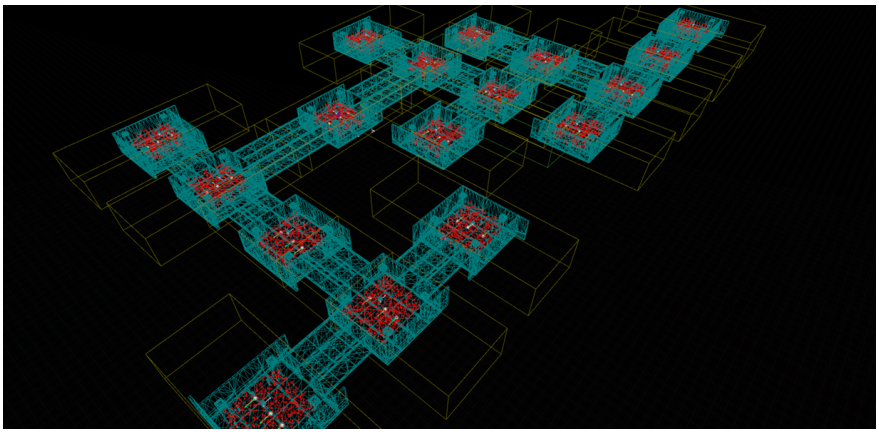




E.4 LitSens in Unreal Engine 4

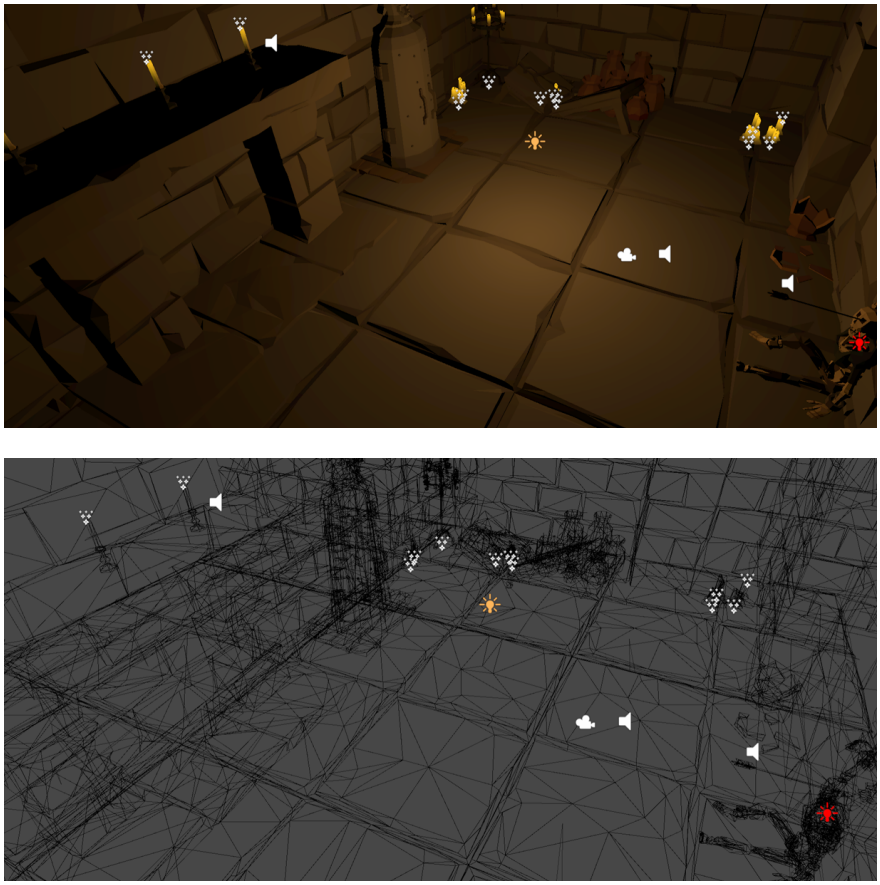
An implementation of the LitSens architecture was built in Unreal Engine 4. Its purpose, as seen in Chapter 7, was to experiment with sound placement in a complex virtual environment (a 3D labyrinth).





E.5 LitSens in Unity

The last implementation of LitSens was built in Unity, and incorporated 6 different levels, each of them related to one of Paul Ekman's basic emotions. Chapter 8 explains how the original architecture was adapted to accept gestures as emotional inputs. A Python thread was running in the background, aside from the engine itself, and was in charge of launching the two different neural network implementations that are detailed in the aforementioned chapter.



E.6 Song of Horror

The application of LPFs to improve recognition of sounds coming from behind was implemented in the commercial video game Song of Horror, from Protocol Games.

APPENDIX F

Appendix F: List of articles supporting this thesis

The following academic articles support the contributions presented in this thesis, and most of their contents are described in the main text:

– M. López Ibáñez, "Bartle Test Applications in Narrative Music Composition for Video Games," in *I Congreso Internacional de Arte, Diseño y Desarrollo de Videojuegos*, (Madrid), pp. 1–13, ESNE, 2015. [65]

– M. López Ibáñez and F. Peinado, "Walking in VR: Measuring Presence and Simulator Sickness in First-Person Virtual Reality Games," in *Proceedings of the 3rd Congreso de la Sociedad Española para las Ciencias del Videojuego*, (Barcelona), pp. 49–60, 2016. [66]

– M. López Ibáñez, N. Álvarez, and F. Peinado, "A Study on an Efficient Spatialisation Technique for Near-Field Sound in Video Games," in *Proceedings of the 4th Congreso de la Sociedad Española para las Ciencias del Videojuego*, (Barcelona), pp. 56–68, 2017. [67]

– M. López Ibáñez, N. Álvarez, and F. Peinado, "LitSens: An Improved Architecture for Adaptive Music Using Text Input and Sentiment Analysis," in *C3Gi 2017*, (Madrid), 2017. [68]

– M. López Ibáñez, N. Álvarez, and F. Peinado, “Towards an Emotion-Driven Adaptive System for Video Game Music,” in *Proceedings of the International Conference on Advances in Computer Entertainment (ACE 2017)*, (London), pp. 360–367, Springer, 2017. [69]

– M. López Ibáñez, N. Álvarez, and F. Peinado, “Assessing the Effect of Adaptive Music on Player Navigation in Virtual Environments,” in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx 2018)* (M. Davies, A. Ferreira, G. Campos, and N. Fonseca, eds.), (Aveiro), pp. 205–212, 2018. [70]

APPENDIX G

Appendix G: Acronym glossary

- 1D: One-dimensional.
- 2D: Two-dimensional.
- 3D: Three-dimensional.
- AI: Artificial intelligence.
- CNN: Convolutional neural network.
- CPU: Central processing unit.
- FEF: Frontal eye field.
- FOV: Field of view.
- FPS: First-person shooter.
- GUI: Graphical user interface.
- HMD: Head-mounted display.
- HRTF: Head-related transfer function.
- LCD: Liquid crystal display.
- LPF: Low-pass filter.

- MIDI: Musical Instrument Digital Interface.
- MLP: Multi-layer perceptron.
- MS: Motion sickness.
- NLTK: Natural Language Toolkit.
- NPC: Non-player character.
- ReLU: Rectified linear unit.
- RAM: Random-access memory.
- RPG: Role-playing game.
- SAM: Self-Assessment Manikin.
- SGD: Stochastic gradient descent.
- SS: Simulator sickness.
- SSQ: Simulator Sickness Questionnaire.
- SUS: Slater-Usoh-Steed Questionnaire.
- TPI: Temple Presence Inventory.
- VR: Virtual reality.
- VST: Virtual Studio Technology.

BIBLIOGRAPHY

[1] S. Adinolf and S. Turkay, "Controlling Your Game Controls: Interface and Customization," in *Proceedings of the 7th International Conference on Games + Learning + Society Conference*, Madison, 2011, pp. 13–22. [Online]. Available: https://dl.acm.org/citation.cfm?id=2206378http://dl.acm.org/citation.cfm?id=2206376.2206378

[2] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/969552/http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=969552

[3] E. Asutay, D. Västfjäll, A. Tajadura-Jiménez, A. Genell, P. Bergman, and M. Kleiner, "Emoacoustics: A study of the psychoacoustical and psychological dimensions of emotional sound design," *AES: Journal of the Audio Engineering Society*, vol. 60, no. 1-2, pp. 21–28, 2012. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16162

[4] P. Balazs, "ARI HRTF Database," 2014. [Online]. Available: https://www.kfs.oeaw.ac.at/

[5] W. Barfield and S. Weghorst, "The sense of presence within virtual environments: A conceptual framework," *Advances in Human Factors Ergonomics*, vol. 19, p. 699, 1993. [Online]. Available: https://scholar.google.es/scholar?q=barfield+sense+of+presence{&}btnG={&}hl=es{&}as{_}sdt=0{%}2C5{#}1

[6] R. Bartle, "Hearts, Clubs, Diamonds, Spades: Players who suit MUDs," *Journal of MUD research*, vol. 6, no. 1, pp. 19–46, 1996.

[7] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," in *Proceedings of the*

International Conference on Learning Representations (ICLR 2015), San Diego, nov 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1511.06448

[8] C. Bateman, R. Lowenhaupt, and L. E. Nacke, "Player Typology in Theory and Practice," in *Proceedings of DiGRA 2011 Conference: Think Design Play*, Hilversum, 2011, pp. 1–24. [Online]. Available: https://hcigames.com/wp-content/uploads/2015/01/Player-Typology-in-Theory-and-Practice.pdf

[9] F. Biocca, C. Harms, and J. K. Burgoon, "Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria," *Presence: Teleoperators and Virtual Environments*, vol. 12, no. 5, pp. 456–480, oct 2003. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/105474603322761270

[10] N. Böttcher, H. P. Martínez, and S. Serafin, "Procedural audio in computer games using motion controllers: An evaluation on the effect and perception," *International Journal of Computer Games Technology*, vol. 2013, no. 6, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2610938

[11] M. Botvinick and J. Cohen, "Rubber hands 'feel' touch that eyes see," *Nature*, 1998. [Online]. Available: http://www.psychology.mcmaster.ca/bennett/psy720/readings/m5/botvinick.pdf

[12] M. L. Bourguet, "Designing and Prototyping Multimodal Commands," in *Interact*, vol. 3, 2003, pp. 717–720.

[13] M. M. Bradley and P. J. Lang, "Measuring emotion: The Self-Assessment Manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[14] B. Cheng and D. M. Titterington, "Neural Networks: A Review from a Statistical Perspective," *Statistical Science*, vol. 9, no. 1, pp. 2–30, 1994. [Online]. Available: https://www.jstor.org/stable/2246275http://dx.doi.org/10.2307/2246275

[15] K. Collins, "An Introduction to Procedural Music in Video Games," *Contemporary Music Review*, vol. 28, no. 1, pp. 5–15, feb 2009. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/07494460802663983

[16] A. Cont, "ANTESCOFO: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music." *Proceedings of the 2008 International Computer Music Conference, Belfast, Northern Ireland*, pp. 33–40, 2008. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00694803/http://articles.ircam.fr/textes/Cont08a/{%}5Cnpapers://616814c5-046b-40e4-b13b-55fce5f60979/Paper/p6954

[17] R. F. Day, C. H. Lin, W. H. Huang, and S. H. Chuang, "Effects of music tempo and task difficulty on multi-attribute decision-making: An eye-tracking approach," *Computers in Human Behavior*, vol. 25, no. 1, pp. 130–143, 2009.

[18] R. Dillon, *On the way to fun*. New York: Taylor & Francis Group, mar 2010. [Online]. Available: https://www.taylorfrancis.com/books/9781439876893

[19] M. Don, C. W. Ponton, J. J. Eggermont, and A. Masuda, "Gender differences in cochlear response time: an explanation for gender amplitude differences in the unmasked auditory brain-stem response." *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2135–48, oct 1993. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.407485http://www.ncbi.nlm.nih.gov/pubmed/8227753

[20] J. L. Dorado and P. A. Figueroa, "Ramps are better than stairs to reduce cybersickness in applications based on a HMD and a Gamepad," in *Proceedings of the IEEE Symposium on 3D User Interfaces 2014 (3DUI 2014)*. Minneapolis: IEEE, 2014, pp. 47–50. [Online]. Available: http://ieeexplore.ieee.org/document/6798841/

[21] J. A. Ehrlich, "Simulator sickness and HMD configurations," *Telemanipulator and Telepresence Technologies*, vol. 3206, no. 4, pp. 170–179, 1997.

[22] J. Eisenberg and W. F. Thompson, "A Matter of Taste: Evaluating Improvised Music," *Creativity Research Journal*, vol. 15, no. 2, pp. 287–296, jul 2003. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10400419.2003.9651421

[23] I. Ekman, "Psychologically Motivated Techniques for Emotional Sound in Computer Games," in *Proceedings of the 3rd Audio Mostly Conference (AM 2008)*, 2008, pp. 20–26. [Online]. Available: https://meaningfulnoise.wordpress.com/psychologically-motivated-techniques-for-emotional-sound-in-computer-games/

[24] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3, pp. 169–200, 1992.

[25] M. Eladhari, R. Nieuwdorp, and M. Fridenfalk, "The Soundtrack of Your Mind: Mind Music - Adaptive Audio for Game Characters," in *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*. Hollywood: ACM, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.9497{&}rep=rep1{&}type=pdf

[26] S. Ellis, "What are virtual environments?" *IEEE Computer Graphics and Applications*, vol. 14, no. 2, pp. 17–22, 1994.

[27] G. Enzner, C. Antweiler, and S. Spors, "Trends in acquisition of individual head-related transfer functions," in *The Technology of Binaural Listening*. Berlin, Heidelberg: Springer, 2013, pp. 57–92. [Online]. Available: http://link.springer.com/10.1007/978-3-642-37762-4{_}3

[28] E. Fedorenko, A. Patel, D. Casasanto, J. Winawer, and E. Gibson, "Structural integration in language and music: Evidence for a shared system," *Memory and Cognition*, vol. 37, no. 1, pp. 1–9, 2009.

[29] A. S. Fernandes and S. K. Feiner, "Combating VR Sickness through Subtle Dynamic Field-Of-View Modification," *3D User Interfaces (3DUI)*, pp. 201–210, 2016. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7460053/

[30] F. Ferri, A. Tajadura-Jiménez, A. Väljamäe, R. Vastano, and M. Costantini, "Emotion-inducing approaching sounds shape the boundaries of multisensory peripersonal space," *Neuropsychologia*, vol. 70, pp. 468–475, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0028393215001062

[31] H. Fletcher and W. A. Munson, "Loudness, Its Definition, Measurement and Calculation," *Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, oct 1933. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6771028

[32] T. A. Galyean, "Guided navigation of virtual environments," in *Proceedings of the 1995 symposium on Interactive 3D graphics - SI3D '95*. New York: ACM Press, 1995, pp. 103–ff. [Online]. Available: http://portal.acm.org/citation.cfm?doid=199404.199421

[33] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995. [Online]. Available: http://www.linux.bucknell.edu/{~}kozick/elec32007/hrtfdoc.pdf

[34] B. Geethanjali, K. Adalarasu, A. Hemapraba, S. P. Kumar, and R. Rajasekeran, "Emotion analysis using SAM (Self-Assessment Manikin) scale," *Biomedical Research*, vol. 28, pp. 18–24, 2017.

[35] M. Grimaldi and P. Cunningham, "Experimenting with music taste prediction by user profiling," in *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval (MIR 2004)*. New York, New York, USA: ACM Press, 2004, pp. 173–180. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1026711.1026740

[36] J. C. Gupta, N., Barreto, A., Joshi, M., & Agudelo, "HRTF Database at FIU DSP Lab," in *2010 IEEE International Conference on Acoustics Speech and*

Signal Processing (ICASSP), Dallas, 2010, pp. 169–172. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/5496084/

[37] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," in *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, vol. 70. JMLR.org, 2017, pp. 1362–1371. [Online]. Available: https://dl.acm.org/citation.cfm?id=3305522http://arxiv.org/abs/1612.01010

[38] M. Handrahan, "Fast Travel Games: Think long and hard on free locomotion in VR," https://www.gamesindustry.biz/articles/2018-06-06-fast-travel-games-think-long-and-hard-on-free-locomotion-in-vr, 2018. [Online]. Available: https://www.gamesindustry.biz/articles/2018-06-06-fast-travel-games-think-long-and-hard-on-free-locomotion-in-vr

[39] Z. Hassani and A. I. Wuryandari, "Music generator with Markov Chain: A case study with Beatme Touchdown," in *Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016*. Bandung: Institute of Electrical and Electronics Engineers, 2017, pp. 179–183. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7849646/

[40] W. Haynes, "Student's t-test," in *Encyclopedia of Systems Biology*. Springer, 2013, pp. 2023–2025. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-1-4419-9863-7{_}1184.pdf

[41] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, aug 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7056445/

[42] D. Hong, T.-h. Lee, Y. Joo, and W.-c. Park, "Real-time Sound Propagation Hardware Accelerator for Immersive Virtual Reality 3D Audio," in *Proceedings of the 21st ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, San Francisco, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?id=3036842

[43] J. Hou, Y. Nam, W. Peng, and K. M. Lee, "Effects of screen size, viewing angle, and players' immersion tendencies on game experience," *Computers in Human Behavior*, vol. 28, no. 2, pp. 617–623, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563211002512zotero://attachment/1060/{%}5Cnhttp://www.sciencedirect.com/science/article/pii/S0747563211002512

[44] I. P. Howard and B. J. Rogers, "Binocular vision and stereopsis," in *Oxford Psychology Series*. Oxford University Press, 1995, vol. 29, pp. 2–4. [On-

line]. Available: https://books.google.es/books?id=I8vqITdETe0C{&}lpg=PA32{&}ots=JhaI-4InNv{&}pg=PA32{&}redir{_}esc=y{#}v=onepage{&}q{&}f=false

[45] ——, *Binocular Vision and Stereopsis*. Oxford University Press, 1995.

[46] C. Jennett, A. Cox, and P. Cairns, "Measuring and defining the experience of immersion in games," *International journal of human-computer studies*, vol. 66, no. 9, pp. 641–661, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1071581908000499

[47] J. Jerald, *The VR book: Human-centered design for virtual reality*. Morgan & Claypool Publishers, 2015. [Online]. Available: https://books.google.es/books?hl=es{&}lr={&}id=ZEBiDwAAQBAJ{&}oi=fnd{&}pg=PR11{&}dq=jason+jerald{&}ots=0yp5DJthX5{&}sig=KT827C4Z1EMD0qRffefnkVkyH8shttp://dl.acm.org/citation.cfm?id=2792790

[48] M. O. Jewell, M. S. Nixon, and A. Prugel-Bennett, "CBS: A concept-based sequencer for soundtrack composition," in *Proceedings of the 3rd International Conference on WEB Delivering of Music (WEDELMUSIC 2003)*, Leeds, 2003, pp. 105–108. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/1233882/

[49] K. Jørgensen, "'What Are These Grunts and Growls Over There?' Computer Game Audio and Player Action," Ph.D. dissertation, 2007. [Online]. Available: http://folk.uib.no/st01206/jorgensen-thesis.pdf

[50] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness," *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, jul 1993. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327108ijap0303{_}3

[51] A. Kiselev, M. Scherlund, A. Kristoffersson, N. Efremova, and A. Loutfi, "Auditory immersion with stereo sound in a mobile robotic telepresence system," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, J. A. Adams and W. Smart, Eds. Portland: ACM, 2015, pp. 55–56. [Online]. Available: https://dl.acm.org/citation.cfm?id=2702034

[52] M. J. Kivland, "The use of music to increase self-esteem in a conduct disordered adolescent," *Journal of Music Therapy*, vol. 23, no. 1, pp. 25–29, 1986.

[53] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch: An Open Source Library for Sentence-Based Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 312–325, 2013.

[54] U. Krcadinac, J. Jovanovic, V. Devedzic, and P. Pasquier, "Textual Affect Communication and Evocation Using Abstract Generative Visuals," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 370–379, 2016.

[55] A. Kulshreshth, K. Pfeil, and J. J. Laviola, "Enhancing the Gaming Experience Using 3D Spatial User Interface Technologies," *IEEE Computer Graphics and Applications*, vol. 38, no. 3, pp. 16–23, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7912169/

[56] O. Lahav, "Improving orientation and mobility skills through virtual environments for people who are blind : Past research and future potential," *International Journal of Child Health and Human Development*, vol. 7, no. 4, pp. 349–355, 2014.

[57] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Systems with Applications*, vol. 41, no. 6, pp. 3041–3046, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417413008518

[58] T. Li, M. Choi, K. Fu, and L. Lin, "Music Sequence Prediction with Mixture Hidden Markov Models," in *4th International Conference on Artificial Intelligence and Applications (AI 2018)*, Dubai, sep 2018, pp. 1–5. [Online]. Available: http://arxiv.org/abs/1809.00842

[59] R. Likert, "A technique for the measurement of attittudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932. [Online]. Available: http://psycnet.apa.org/psycinfo/1933-01885-001http://www.voteview.com/pdf/Likert{_}1932.pdf{%}5Cnhttp://psycnet.apa.org/psycinfo/1933-01885-001

[60] R. P. Lippmann, "Review of Neural Networks for Speech Recognition," *Neural Computation*, vol. 1, no. 1, pp. 1–38, mar 2008. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/neco.1989.1.1.1

[61] S. Little and E. Mena, "GameTrack European Summary Data," ISFE, Tech. Rep., 2018. [Online]. Available: https://www.isfe.eu/sites/isfe.eu/files/gametrack{_}european{_}summary{_}data{_}2018{_}q3.pdf

[62] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. Boston, MA: Springer US, 2012, vol. 9781461432, pp. 415–463. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-3223-4{_}13

[63] S. Livingstone and A. Brown, "Dynamic response: real-time adaptation for music emotion," in *Second Australasian Conference on Interactive Entertainment*. Sydney:

Creativity & Cognition Studios Press, 2005, pp. 105–111. [Online]. Available: http://dl.acm.org/citation.cfm?id=1109196http://portal.acm.org/citation.cfm?id=1109196

[64] M. Lombard, T. Ditton, and L. Weinstein, "Measuring Presence: The Temple Presence Inventory," in *Proceedings of the 12th Annual International Workshop on Presence*, 2009, pp. 1–15. [Online]. Available: http://astro.temple.edu/{~}tuc16417/papers/Lombard{_}et{_}al.pdf

[65] M. López Ibáñez, "Bartle Test Applications in Narrative Music Composition for Video Games," in *I Congreso Internacional de Arte, Diseño y Desarrollo de Videojuegos*. Madrid: ESNE, 2015, pp. 1–13.

[66] M. López Ibáñez and F. Peinado, "Walking in VR: Measuring Presence and Simulator Sickness in First-Person Virtual Reality Games," in *Proceedings of the 3rd Congreso de la Sociedad Española para las Ciencias del Videojuego*, Barcelona, 2016, pp. 49–60.

[67] M. López Ibáñez, N. Álvarez, and F. Peinado, "A Study on an Efficient Spatialisation Technique for Near-Field Sound in Video Games," in *Proceedings of the 4th Congreso de la Sociedad Española para las Ciencias del Videojuego*, Barcelona, 2017, pp. 56–68. [Online]. Available: http://ceur-ws.org/Vol-1957/

[68] ——, "LitSens: An Improved Architecture for Adaptive Music Using Text Input and Sentiment Analysis," in *Proceedings of the C3Gi Conference 2017*, Madrid, 2017.

[69] ——, "Towards an Emotion-Driven Adaptive System for Video Game Music," in *Proceedings of the International Conference on Advances in Computer Entertainment (ACE 2017)*. London: Springer, 2017, pp. 360–367.

[70] ——, "Assessing the Effect of Adaptive Music on Player Navigation in Virtual Environments," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx 2018)*, M. Davies, A. Ferreira, G. Campos, and N. Fonseca, Eds., Aveiro, 2018, pp. 205–212.

[71] M. Luhtala, M. Turunen, J. Hakulinen, and T. Keskinen, "AIE-studio - A pragmatist aesthetic approach for procedural sound design," in *Proceedings of the 8th Audio Mostly Conference (AM 2013)*, 2013, pp. 7–14. [Online]. Available: http://dl.acm.org/citation.cfm?id=2544124http://dl.acm.org/citation.cfm?id=2544114.2544124{%}5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84898834763{&}partnerID=tZOtx3y1

[72] M. A. Martínez, "Storyworld Possible Selves and the Phenomenon of Narrative Immersion: Testing a New Theoretical Construct," *Narrative*, vol. 22, no. 1, pp. 110–

131, 2014. [Online]. Available: https://muse.jhu.edu/article/536493/summaryhttp://muse.jhu.edu/content/crossref/journals/narrative/v022/22.1.martinez.html

[73] J. H. McDermott and A. J. Oxenham, "Music perception, pitch, and the auditory system," *Current Opinion in Neurobiology*, vol. 18, no. 4, pp. 452–463, 2008.

[74] A. Mehrabian and J. A. Russell, *An approach to environmental psychology*. Cambridge: The MIT Press, 1974. [Online]. Available: http://psycnet.apa.org/record/1974-22049-000http://web.mit.edu/index.html

[75] M. Mera, "Towards 3-D sound: Spatial presence and the space vacuum," in *The Palgrave Handbook of Sound Design and Music in Screen Media: Integrated Soundtracks*. London: Palgrave Macmillan UK, 2016, pp. 91–111. [Online]. Available: http://link.springer.com/10.1057/978-1-137-51680-0{_}7

[76] O. Merhi and T. A., "Motion Sickness, Video Games, and Head-Mounted Displays," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 23. Chicago: SAGE Publications, sep. 2004, pp. 2618–2622. [Online]. Available: http://pro.sagepub.com/content/48/23/2618.abstract

[77] D. Mestre, P. Fuchs, A. Berthoz, and J. L. Vercher, "Immersion et présence," *Le traité de la réalité virtuelle. Paris: Ecole des Mines de Paris*, pp. 309–338, 2006.

[78] D. Milam and M. S. El Nasr, "Design Patterns to Guide Player Movement in 3D Games," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games - Sandbox '10*, vol. 1, no. 212. New York: ACM Press, 2010, pp. 37–42. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1836135.1836141

[79] W. Min, B. Mott, J. Rowe, B. Liu, J. Lester, and N. Carolina, "Player Goal Recognition in Open-World Digital Games with Long Short-Term Memory Networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, New York, 2016, pp. 2590–2596. [Online]. Available: https://www.ijcai.org/Proceedings/16/Papers/368.pdf

[80] W. Mitchell, "Journal of Visual Culture," *Journal of Visual Culture*, vol. 1, no. 2, pp. 165–181, 2002. [Online]. Available: http://journals.sagepub.com/doi/abs/10.1177/147041290200100202

[81] D. Monteiro, H. N. Liang, W. Xu, M. Brucker, V. Nanjappan, and Y. Yue, "Evaluating enjoyment, presence, and emulator sickness in VR games based on first- and third-person viewing perspectives," in *Computer Animation and Virtual Worlds*, vol. 29, no. 3-4, may 2018. [Online]. Available: http://doi.wiley.com/10.1002/cav.1830

[82] M. Morimoto and Y. Ando, "On the Simulation of Sound Localization," *Journal of the Acoustical Society of Japan*, vol. 1, pp. 167–174, 1980. [Online]. Available: https://www.jstage.jst.go.jp/article/ast1980/1/3/1{_}3{_}167/{_}article/-char/ja/

[83] A. Nayebi and M. Vitelli, "GRUV: Algorithmic Music Generation using Recurrent Neural Networks," Tech. Rep., 2015. [Online]. Available: https://web.stanford.edu/{~}anayebi/projects/CS{_}224D{_}Final{_}Project{_}Writeup.pdf

[84] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Sciences Education*, vol. 15, no. 5, pp. 625–632, 2010.

[85] L. C. Ou, M. R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. Part I: Colour emotions for single colours," *Color Research and Application*, vol. 29, no. 3, pp. 232–240, 2004.

[86] F. Pachet, "The Continuator: Musical Interaction With Style," *Journal of New Music Research*, vol. 32, no. 3, pp. 333–341, sep 2003. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1076/jnmr.32.3.333.16861

[87] F. Pachet and P. Roy, "Markov constraints: Steerable generation of Markov sequences," *Constraints*, vol. 16, no. 2, pp. 148–172, apr 2011. [Online]. Available: http://link.springer.com/10.1007/s10601-010-9101-4

[88] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 2010, pp. 1320–1326. [Online]. Available: http://incc-tps.googlecode.com/svn/trunk/TPFinal/bibliografia/PakandParoubek(2010).TwitterasaCorpusforSentimentAnalysisandOpinionMining.pdf

[89] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Stroudsburg: Association for Computational Linguistics, 2002, pp. 79–86. [Online]. Available: https://dl.acm.org/citation.cfm?id=1118704

[90] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," in *Foundations and Trends in Information Retrieval*, 2008, vol. 2, no. 1–2. [Online]. Available: http://www.nowpublishers.com/article/Details/INR-011

[91] J. Plouzeau, D. Paillot, J. Chardonnet, and F. Merienne, "Effect of proprioceptive vibrations on simulator sickness during navigation task in virtual environment," in *Proceedings of the International Conference on Artificial Reality and Telexistence*

Eurographics Symposium on Virtual Environments, 2015. [Online]. Available: http://sam.ensam.eu/handle/10985/10422

[92] I. Poole, "What is a Chebyshev RF Filter - the Basics | Electronics Notes," https://www.electronics-notes.com/articles/radio/rf-filters/what-is-chebychev-filter-basics.php, 2012. [Online]. Available: https://www.electronics-notes.com/articles/radio/rf-filters/what-is-chebychev-filter-basics.php

[93] R. Ramadan and Y. Widyani, "Game development life cycle guidelines," in *2013 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2013*. IEEE, sep 2013, pp. 95–100. [Online]. Available: http://ieeexplore.ieee.org/document/6761558/

[94] J. Redondo, I. Fraga, I. Padrón, and A. Piñeiro, "Affective ratings of sound stimuli," *Behavior Research Methods*, vol. 40, no. 3, pp. 784–790, aug 2008. [Online]. Available: http://www.springerlink.com/index/10.3758/BRM.40.3.784

[95] R. J. Ritsma, "Frequencies Dominant in the Perception of the Pitch of Complex Sounds," *The Journal of the Acoustical Society of America*, vol. 42, no. 1, pp. 191–198, jul 1967. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1910550

[96] A. Rolnick and R. E. Lubow, "Why is the driver rarely motion sick? The role of controllability in motion sickness," *Ergonomics*, vol. 34, no. 7, pp. 867–879, 1991.

[97] M. Rosenblatt, "A Central Limit Theorem and a Strong Mixing Condition," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, no. 1, jan 1956, pp. 43–47. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16589813http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC534230

[98] D. Roth, "Procedurally Generated Classical Music: Baroque Dances v2," http://devinrothmusic.com/baroquedances/v2.php. [Online]. Available: http://devinrothmusic.com/baroquedances/v2.php

[99] ——, "Generative Music: Jazz Cycle," https://github.com/devinroth/GenerativeMusic/blob/master/JazzCycle.playground/Contents.swift, 2016.

[100] M. Scharkow, R. Festl, J. Vogelgesang, and T. Quandt, "Beyond the "core-gamer": Genre preferences and gratifications in computer games," *Computers in Human Behavior*, vol. 44, pp. 293–298, 2015.

[101] S. M. Schneider, C. K. Kisby, and E. P. Flint, "Effect of virtual reality on time perception in patients receiving chemotherapy." *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*, vol. 19,

no. 4, pp. 555–64, apr 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3673561{&}tool=pmcentrez{&}rendertype=abstract

[102] S. R. Serge and A. D. Moss, "Simulator sickness and the oculus rift: A first look," in *Proceedings of the Human Factors and Ergonomics Society*. Los Angeles: SAGE Publications, 2015, pp. 761–765.

[103] T. B. Sheridan, "Musings on telepresence and virtual presence," *Presence: Teleoperators and virtual environments*, vol. 1, no. 1, pp. 120–126, 1992.

[104] W. Sherman and A. Craig, *Understanding virtual reality: Interface, application, and design*. San Francisco: Elsevier Science, 2003. [Online]. Available: https://books.google.es/books?hl=es{&}lr={&}id=b3OJpAMQikAC{&}oi=fnd{&}pg=PP1{&}dq=understanding+virtual+reality{&}ots=3DKMjsyNOl{&}sig=EPfhSkMH-E1uJztUExrBpu4ZR2A

[105] M. Slater and S. Wilbur, "A framework for immersive virtual environments (FIVE)," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, 1997. [Online]. Available: https://www.mitpressjournals.org/doi/abs/10.1162/pres.1997.6.6.603

[106] A. Slater, M., Usoh, M., & Steed, "Depth of Presence in Virtual Environments," *Presence: Teleoperators and Virtual Environments*, vol. 3, pp. 130–144, 1994.

[107] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 1999. [Online]. Available: http://www.dspguide.com

[108] K. Squire, "Open-Ended Video Games : A Model for Developing Learning for the Interactive Age," in *The Ecology of Games*, K. Salen Tekinbaş, Ed. MIT Press, 2008, pp. 167–198. [Online]. Available: http://website.education.wisc.edu/kdsquire/manuscripts/squire-open-ended-games-macarthur-salen.pdfhttp://www.mitpressjournals.org/doi/abs/10.1162/dmal.9780262693646.167

[109] W. Strank, "The Legacy of iMuse: Interactive Video Game Music in the 1990s," in *Music and Game: Perspectives on a popular alliance*. Wiesbaden: Springer, 2013, pp. 81–91. [Online]. Available: http://link.springer.com/10.1007/978-3-531-18913-0{_}4http://link.springer.com/10.1007/978-3-531-18913-0

[110] Q. Sun, A. Kaufman, A. Patney, L.-Y. Wei, O. Shapira, J. Lu, P. Asente, S. Zhu, M. Mcguire, and D. Luebke, "Towards virtual reality infinite walking," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–13, 2018. [Online]. Available: https://dl.acm.org/citation.cfm?id=3201294http://dl.acm.org/citation.cfm?doid=3197517.3201294

[111] J. Treleaven, J. Battershill, D. Cole, C. Fadelli, S. Freestone, K. Lang, and H. Sarig-Bahat, "Simulator sickness incidence and susceptibility during neck motion-controlled virtual reality tasks," *Virtual Reality*, vol. 19, no. 3-4, pp. 267–275, nov 2015. [Online]. Available: http://link.springer.com/10.1007/s10055-015-0266-4

[112] A. C. Tsoi and A. D. Back, "Locally Recurrent Globally Feedforward Networks: A Critical Review of Architectures," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 229–239, 1994. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/279187/

[113] D. Tsonos and G. Kouroupetroglou, "A Methodology for the Extraction of Reader's Emotional State Triggered from Text Typography," in *Tools in Artificial Intelligence*. InTech, aug 2008.

[114] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Philadelphia: Association for Computational Linguistics, 2002, pp. 417–424. [Online]. Available: https://dl.acm.org/citation.cfm?id=1073153

[115] M. Usoh, K. Arthur, M. C. Whitton, R. Bastos, A. Steed, M. Slater, and F. P. Brooks, "Walking > walking-in-place > flying, in virtual environments," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99*. New York: ACM Press, jul 1999, pp. 359–364. [Online]. Available: http://dl.acm.org/citation.cfm?id=311535.311589

[116] Y. Volcani and D. Fogel, "System and method for determining and controlling the impact of text," U.S. Patent No. 7,136,877. Washington, DC: U.S. Patent and Trademark Office, 2006. [Online]. Available: https://patents.google.com/patent/US7136877B2/en

[117] O. Warusfel, "LISTEN HRTF database," http://recherche.ircam.fr/equipes/salles/listen/, 2002. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[118] T. Wijman, "Mobile Revenues Account for More Than 50% of the Global Games Market as It Reaches $137.9 Billion in 2018," https://newzoo.com/insights/articles/global-games-market-reaches-137-9-billion-in-2018-mobile-games-take-half/, 2018. [Online]. Available: https://newzoo.com/insights/articles/global-games-market-reaches-137-9-billion-in-2018-mobile-games-take-half/

[119] A. Williams and F. J. Taylor, *Electronic Filter Design Handbook*. McGraw-Hill, 1995. [Online]. Available: http://cds.cern.ch/record/327913

[120] S. Wolfson and G. Case, "Effects of sound and colour on responses to a computer game," *Interacting with Computers*, vol. 13, no. 2, pp. 183–192, 2000.

[121] R. Yalch and E. Spangenberg, "Effects of Store Music on Shopping Behavior," *Journal of Services Marketing*, vol. 7, no. 2, pp. 55–63, 1990.

[122] H. Yanagisawa, T. Murakami, S. Noguchi, K. Ohtomi, and R. Hosaka, "Quantification Method of Diverse Kansei Quality for Emotional Design: Application of Product Sound Design," in *Proceedings of the International Design Engineering Technical Conference (DETC 2007)*, vol. 7, Las Vegas, 2007, pp. 461–470. [Online]. Available: http://proceedings.asmedigitalcollection.asme.org/pdfaccess.ashx?ResourceID=5285016{&}PDFSource=13

[123] M. Yang and P. Sheng, "Sound Absorption Structures: From Porous Media to Acoustic Metamaterials," *Annual Review of Materials Research*, vol. 47, no. 1, pp. 83–114, jul 2017. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-matsci-070616-124032

[124] S. Yao, Y. Zhao, H. Shao, A. Zhang, C. Zhang, S. Li, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*, vol. 1, no. 4, Perth, 2017, pp. 351–360. [Online]. Available: https://dl.acm.org/citation.cfm?id=3052577

[125] N. Yee, "The demographics, motivations, and derived experiences of users of massively multi-user online graphical environments," *Presence: Teleoperators and Virtual Environments*, vol. 15, no. 3, pp. 309–329, jun 2006. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/pres.15.3.309

[126] A. C. Younkin and P. J. Corriveau, "Determining the amount of audio-video synchronization errors perceptible to the average end-user," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 623–627, sep 2008. [Online]. Available: http://ieeexplore.ieee.org/document/4599253/

[127] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement." *Emotion*, vol. 8, no. 4, pp. 494–521, 2008. [Online]. Available: http://psycnet.apa.org/journals/emo/8/4/494/http://doi.apa.org/getdoi.cfm?doi=10.1037/1528-3542.8.4.494

[128] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," *Lecture Notes in Computer Science*, vol. 8485 LNCS, pp. 298–310, 2014. [Online]. Available: http://link.springer.com/10.1007/978-3-319-08010-9{_}33