CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF ELECTRICAL ENGINEERING
DEPARTMENT OF RADIOELECTRONICS

DIPLOMA THESIS

VIRTUAL ACOUSTIC SPACE TEST USING HMD

VIKTOR JAROLÍMEK

Supervisor: Ing. František Rund, Ph.D.

Study Programme: Electronics and Communication
Specialization: Media and Signal Processing

Prague 2019

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Jarolímek Viktor** Personal ID number: **434669**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Radioelectronics**

Study program: **Electronics and Communications**

Branch of study: **Media and Signal Processing**

II. Master’s thesis details

Master’s thesis title in English:

Virtual Acoustic Space Test using HMD

Master’s thesis title in Czech:

Testování virtuálního akustického prostoru s využitím HMD

Guidelines:

Get familiar with theory on virtual acoustic space (VAS). Do a background research on the topic, focusing on audio quality testing in VAS. Learn about creating content for VR head mounted displays (HMD), choose a suitable platform, fit for testing in virtual acoustic space. Design and implement an environment for audio quality testing in created VAS using HMD. Prepare and verify pilot tests, designed for changes perceived audio location in regard to used HRTF library, AV distractors etc.

Bibliography / sources:

[1] CARLILE, Simon (ed.). Virtual auditory space: Generation and applications. Springer, 1996.
[2] PERNAUX, Jean-Marie; EMERIT, Marc; NICOL, Rozenn. Perceptual evaluation of binaural sound synthesis: the problem of reporting localization judgments. In: Audio Engineering Society Convention 114. Audio Engineering Society, 2003.

Name and workplace of master’s thesis supervisor:

Ing. František Rund, Ph.D., Department of Radioelectronics, FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **14.02.2019** Deadline for master's thesis submission: **24.05.2019**

Assignment valid until: **20.09.2020**

Ing. František Rund, Ph.D. Supervisor’s signature	prof. Mgr. Petr Páta, Ph.D. Head of department’s signature	prof. Ing. Pavel Ripka, CSc. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt Student’s signature

AUTHOR STATEMENT FOR UNDERGRADUATE THESIS

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university thesis.

PROHLÁŠENÍ AUTORA PRÁCE

Prohlašuji, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

.... 2019 in Prague

....

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisor, Ing. František Rund Ph.D., who has helped me throughout the entire project. I would also like to thank him for his support, patience and guidance he provided me at any time I needed it, because without his ideas and vision, this thesis would not have reached its full potential. I would also like to thank all the kind and wise people of StackOverflow website, who have, over the years, answered so many unearthly and perplexing questions, which allowed me to solve many obstacles I encountered while working on this thesis.

This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/190/OHK3/3T/13.

ABSTRACT

This thesis presents the process of design, development and testing of an application, used for testing virtual acoustic space. The paper describes the entire process of choosing the right parameters to monitor, based on background research into previously developed tests and applications, and selecting the most suitable environment and approach towards the project. A general summary of reviewed projects and papers is also included, along with the full course of creating an easy to use and flexible tool to test parameters of virtual acoustic space and HRTF quality assessment. First test design, outcoming conclusions and final test design, featuring various test cases and HRTF exchange possibility, are described in detail. Subjective tests on ten different subjects were conducted using the developed application and yielded results confirming its full functionality and presented ideas for future studies.

Keywords: virtual reality, audio quality, HRTF, Unity, head mounted display, HMD, virtual acoustic space

ABSTRAKT

Tato diplomová práce pojednává o průběhu návrhu, vývoje a testování aplikace využitelné pro testování virtuálního akustického prostoru. Práce popisuje celý proces výběru správných parametrů ke sledování, vycházející z rešerše na již vytvořené experimenty a aplikace, výběru nejvhodnějšího prostředí a celkového přístupu k celému projektu. Zahrnuto je také obecné shrnutí nastudovaných projektů a článků, spolu s popisem celého průběhu vytváření nástroje, který by byl jednoduchý a flexibilní pro testování parametrů virtuálního akustického prostoru a kvality HRTF. Detailně jsou popsány první návrh testu, z něho plynoucí poznatky a finální návrh testů s různorodým spektrem dílčích částí experimentu a možností výměny HRTF. Deset dobrovolníků podstoupilo subjektivní testy za použití vyvinuté aplikace a získané výsledky potvrzují její plnou funkcionalitu a představují nápady pro budoucí práci.

Klíčová slova: virtuální realita, kvalita zvuku, HRTF, Unity, brýle pro virtuální realitu, HMD, virtuální akustický prostor

Contents

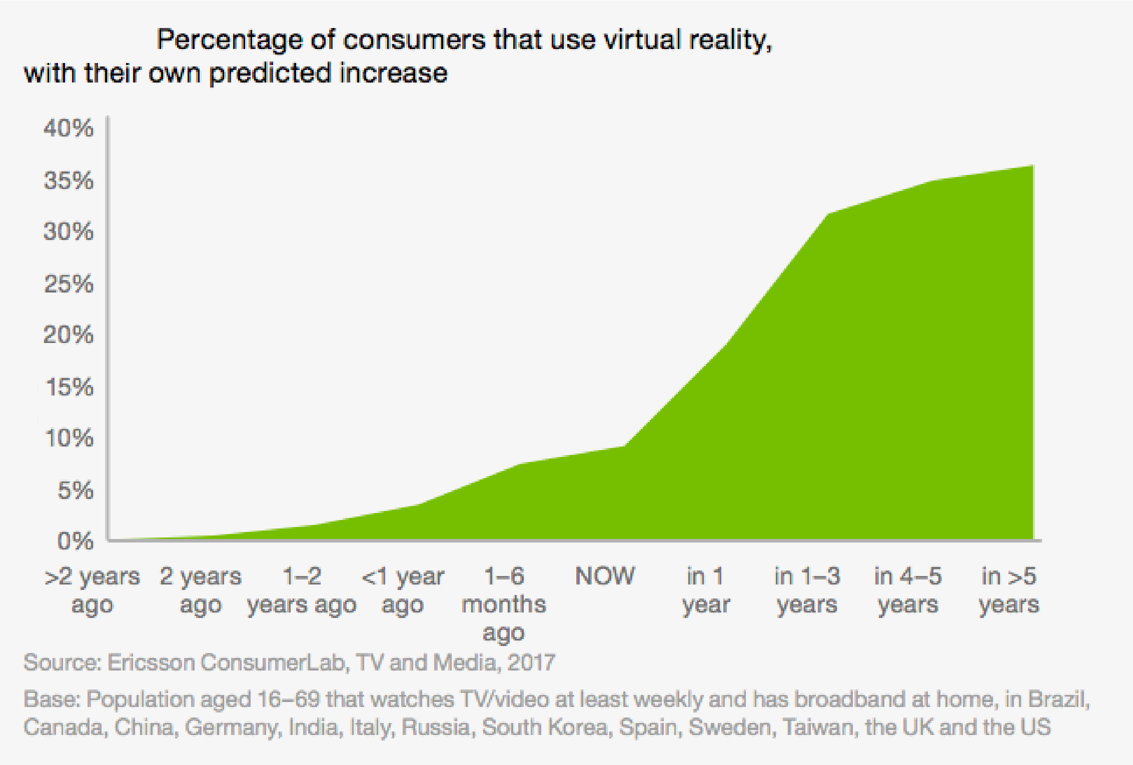
GLOSSARY

AES	–	Audio Engineering Society (page 10)
AI	–	Artificial Intelligence (page 8)
AV	–	Audio-Visual (page 9)
DoF	–	Degrees of Freedom (page 24)
GUI	–	Graphical User Interface (page 22)
HMD	–	Head Mounted Display (page 5)
HRIR	–	Head Related Impulse Response (page 14)
HRTF	–	Head Related Transfer Function (page 6)
IID	–	Interaural Intensity Difference (page 13)
ILD	–	Interaural Level Difference (page 12)
IR	–	Impulse Response (page 14)
ITD	–	Interaural Time Difference (page 12)
JND	–	Just Noticeable Difference (page 15)
MAA	–	Minimal Audible Angle (page 15)
MRTK	–	Windows Mixed Reality Toolkit (page 19)
SOFA	–	Spatially Oriented Format for Acoustics (page 10)
VAS	–	Virtual Acoustic Space (page 6)
VR	–	Virtual Reality (page 5)

1. INTRODUCTION

1.1. MOTIVATION

Virtual reality (VR), in the 3D, inclusive way, is becoming increasingly integrated within our lives these days and its popularity is expected to rise even more over the following years [1] (an estimate can be seen in Figure 1 below this paragraph). Not only there is an abundant number of games for various platforms (Xbox, PlayStation, Head Mounted Displays (HMDs) like HTC Vive and Oculus and many more), but there are also various applications for industry, learning, teaching and other areas. One of the key aspects for each application is sound. Among other reasons, virtual reality was built and intended for maximum immersion. In other words, people should not be able to distinguish between reality and virtual reality or at least be able to fully comprehend the environment, built in the virtual world. For that idea to fully work, all of our senses need to be included within the virtual experience. Without realizing it, all of our senses come together when we interact during every minute of our daily life. When one of these senses is off, our brain is quickly alarmed and that is when we start to question the environment, we are in. For example, that is also the reason, why we feel so comfortable, when the controllers are exactly where we see them in VR, when we reach for them. Virtual reality (in the sense of immersive VR) by itself is always based in vision, that is why we either need to use head mounted displays (HMDs), like HTC Vive, Oculus, Acer and many others, or cave systems, where the screens all around us. The other important sense for us to make it feel more real, is sound, and to be more precise, a spatial sound. For the process of learning, entertainment and a correct VR based behavior as a whole, sound that is used in these applications needs to be as precise as possible. That also includes the quality of the sound (sampling frequency, bit depth etc.), but mainly the ability for the user to localize the sound to the, preferably, exact position as one would be in real life.



Percentage of consumers that use virtual reality, with their own predicted increase

Source: Ericsson ConsumerLab, TV and Media, 2017

Base: Population aged 16–69 that watches TV/video at least weekly and has broadband at home, in Brazil, Canada, China, Germany, India, Italy, Russia, South Korea, Spain, Sweden, Taiwan, the UK and the US

Figure 1: The increase of popularity of VR throughout the years [1]

The goal of this work was to create an application that would allow us and others, to some extent, measure this ability. How precise is the system rendering the location of the currently playing audio source, while also including the imperfections of human auditory system, allowing given application to have some error without disturbing the user or deviating from its purpose. Firstly, we wanted to be able to measure the quality of the sound location and its back rendering – testing general and specific head related transfer functions (HRTFs – *2.4.1. HRTF, ITD, ILD*). Secondly, our goal was to also determine, how big of an error is still acceptable for the user, before it becomes disturbing to the VR experience.

1.2. ASSIGNMENT

The purpose of this thesis was to develop a way to test the audio quality in virtual reality. And one of the main criteria was to connect the auditory stimulus with a visual one. When we involve only one of our senses, we tend to perceive in a different way than if we are introduced to multiple stimuli. There are many ways to test audio quality and many parameters that can be measured, so the first part of the assignment was to do a research on current, previous and regular tests in order for this paper to have a real value and not to only recreate someone's work. In regards to the background research a more specific test was to be designed and only some parameters to be chosen. Also, a working device had to be picked and purchased, a programming environment/engine had to be selected and the general architecture of the test had to be proposed. The final goal was to create an application that would allow us to test HRTFs and certain parameters that would partially determine the quality of audio in virtual acoustic space (VAS). This application was to be tested using subjective tests, measuring the selected parameters and determining the overall functionality of the application.

1.3. THESIS STRUCTURE

First part of the thesis consists of an introductory chapter (*2. Theoretical introduction*). Going into detail on various problematics and research prior making specific decisions. Topics of virtual reality, human perception, development for virtual reality and virtual acoustic space are discussed. Brief overview of audio quality measurements, subjective testing, virtual reality display devices is given and at the end of this section a thorough background research into virtual acoustic space audio quality is analyzed.

Following that, a section describing initial test design (*3. Subjective test design*) presents decisions based on data from the theoretical introduction chapter. Moreover, initial test creation is described and conclusions are drawn from its functionality. Based on the findings from the first stage, a new test architecture and test flow is introduced and additional assets are parametrized (*3.3. Final test design*). This section also features description of the final testing phase (*3.4. Final testing*).

In the second to last section, results are processed and a discussion is led regarding their outcomes (*4. Results and discussion*). A short evaluation of the selected devices and platform is included, along with application assessment. Finally, a conclusion and a summary of the entire work is presented (*5. Conclusion*).

2. THEORETICAL INTRODUCTION

2.1. VIRTUAL REALITY

Virtual reality (VR) may seem like a simple concept, but the definition is not as straightforward as people may think [2]. Today, when we say VR, we usually imagine the virtual reality headsets and the all virtual environment it provides. Nowadays that is what usually people mean, but in general, virtual reality is exactly what it sounds like. A fully virtual world that can have many forms. One of the most successful ones was World of Warcraft [3]. A game that is run on computer and features the possibility to create and improve your character in a multiplayer gameplay environment. The multiplayer delivery might be a key part of it, because that it what substitutes the reality, the feeling of belonging in the game with others. And of course, it is not the only one. However, nowadays, VR is truly meant as the fully virtual environment that lets us interact with objects around us that are rendered via the display device. The more correct label would be immersive VR [4] as immersion is the crucial aspect of this technology.

Immersion is, by definition, a feeling of involvement. In case of VR, the involvement within a virtual environment, designed by developers. The user has the possibility to interact with the world around him – the combination of this interaction and immersion is called telepresence. In the perfect scenario, the user completely forgets he is in a virtual world – the world becomes indistinguishable from the world we live in. According to [5], immersion is made up of two components – depth of information and breadth of information. Depth of information is made up of anything from resolution of the display, quality of graphics, audio quality to pretty much any data. On the other hand, breadth of information can be defined as number of sensations that are simulated simultaneously. At the moment, only audio and video are senses that are commonly researched and used. There are, however, systems that simulate the sense of touch – haptic systems. And as the newest technology, some startups have started developing devices that would simulate smells and fragrances along with some additional sensory information, e.g. wind, water mist and heat [6]. There are of course additional aspects to immersion. Even the size of the environment, or the scale of possible interactions with the objects and so on.

2.2. HUMAN PERCEPTION IN VR

As partly mentioned in the beginning, we are not built to function within virtual reality. There are so many aspects of our everyday lives that are automatic for us, but may easily disrupt our perception in VR. And are also hard to credibly reproduce to ensure the highest immersion.

One of the best examples is movement within the virtual world. When we walk around in real life, our body produces a certain continuous motion and our body and head move along with every step. We do not perceive this in a way that our vision would blur or sway, we do not even notice it. Similar to breathing. This natural motion is very hard to imitate and as our brain is used to compensating for it, the lack of it instantly induces deep discomfort. In regular desktop computer/gaming console applications, we move using joysticks, arrows and similar. In VR, this would be impossible and the continuous motion would be in dissent with our usual perception and would result in nausea. For the brave of heart, it can be easily simulated when the HMD loses tracking and the vision starts drifting. For that reason, movement around the virtual environment is solved using teleportation and rotations with fade in/fade out effects and is always presented as “jumps” between two positions.

Sound is as much a part of the virtual world as visual content and to create the feeling of immersion, we need to perceive the virtual sound spatially as we would outside of VR. It is however very different whether we only perceive spatial sound alone, or whether it is inside a virtual world. This is also a premise derived from the assignment and the goal of this thesis as the important factor is connecting audio and visual stimuli. Tests, solely focused on audio, with no visual data to accompany it, would definitely have different results to those, where a visual stimulus is present at the same time as sound. With pure audio testing, people are entirely focused on the task at hand, and are therefore able to spot less significant differences. My belief is that when a visual stimulus is introduced, senses come together and are willing to interpolate for or overlook certain imperfections (this will also be mentioned in some studies in the *2.8. Background research on VAS audio quality and HRTF based measurements* chapter). Psychoacoustics field plays a vital role when it comes to perception in virtual reality, as we are also less likely to perceive errors in sound when the source is based in its “natural environment”, or can be explained by something that is happening in the scene. These are so called diegetic sounds [7]. Same concept as in real life, a machine sound is much more alarming to us when heard in nature than in a city. Moreover, the replayed audio sample cannot be irritating as it would distort the illusion. Another and entirely different aspects are so called distractors, which can be presented in many forms, and in my case, the most important would be audio and visual distractors. These might be present as for example secondary audio sources, ambient noise – for audio – or similarly looking objects and a high number of items – for video.

One of the key aspects of our hearing is experience. There are directions and elevations our hearing is not able to safely a precisely recognize and localize (see *2.4.2. Cone of confusion*). Or at least it would not be if we heard it for the first time in our life. We are, however, able to react to these sounds because of our lifelong experience. It can be imagined as an artificial intelligence (AI) learning algorithm. As we grow up, we are faced with various sounds and situations, where we need to, or want to, localize the source of the sound/music and the knowledge is stored and used in the future.

These ambiguities can also be solved using head movement and rotation, which we do subconsciously. Our body is driven by reflexes and some of these take part and take over during localization process. That is why we tend to turn, when we hear a very loud or an unknown sound. That is both an evolutional trait (to see and asses possible danger) as well as a way to closely inspect the direction of the sound and localize the source.

As seen from these examples, our brain uses every sensory resource available and constantly evaluates every situation we find ourselves in. For that reason, it is a never-ending task, trying to recreate the real world virtually. And integrating a believable and possible connection between video and audio stimuli might be one of the keys.

2.3. CHOOSING A WORKING PLATFORM

There are many viable options, each possible, but some better suited for different types of applications and desired outputs. One of the options would be using only an integrated development environment and create everything from scratch. That might present an unnecessary obstruction, for the assignment of this extent, as the time consumption would be enormous. However, it would also provide absolute control over the application's behavior and aspects. Apart from this, there is a vast selection of different platforms that are designed for programmers in order to save them some precious time. Engines meant for development for non-programmers, programmers, people who prefer designing logic over graphics or exactly the other way around and many more. Differences range from the availability (paid/free), support, forums, up to the focused output platform (computer, console, 2D, 3D, VR and so on).

2.3.1. GAME ENGINES AS A SOLUTION

In the spirit of efficiency, both programming and time wise, game engines present a suitable platform for creating the VR content of the test. Considering the primary goal was to create an audio-visual (AV) test, game engine would allow creating graphics and VR environment quickly and efficiently with a lot of versatility with the lowest time consumption possible (in comparison to creating a graphic content via code or 3D modeling software). There are various possibilities available for students and beginning developers, but the most famous and widespread ones are Unity [8] and Unreal Engine [9]. Both of these have a long tradition and many supporters and forums. Moreover, they both serve as the backbone of some of world-renowned projects throughout a wide spectrum of many scientific and application fields.

Various aspects have to be considered though – programming language, forums, production company support, plugins, free and available content, tutorials, understandability of the environment, possible workflow and more. Many of these parameters are more or less similar in both of these game engines. Nevertheless, one of the dissimilar aspects is the programming language and the approach to programming in general. After going through the community forums and various discussion, the consensus appears to have a more or less direct result. Unreal Engine is powered by C++ and Blueprints, which make work for beginner programmers a lot easier. In other words, it is easier, for beginner developer, to create a polished, good looking game/application, without the need of in-depth programming. It allows working with prepared blueprints and creating schemes (game logic) with beginner designer knowledge. However, C++ is a high-level programming language and is considered to be one of the most complex and versatile languages, but also a one of the hardest. By contrast, Unity runs on C# (and alternatively JavaScript), which, despite being a high-level programming language as well, is easier to learn and still offers a complete control over all aspects of any application written in it. Some argue, that Unreal Engine might be simpler in the beginning but presents more obstacles further on. Others say that even though Unity possibly has a smaller learning curve, some things do not clearly add up and lack consistency. In the online community, the "battle" between Unity and Unreal Engine, has no clear winner, as some people prefer one over the other, but both engines still undoubtedly serve its purpose and present powerful tools.

Apart from that, Unity and Unreal Engine offer various VR plugins and support every major HMD developer – Steam VR, Mixed Virtual Reality and Oculus VR. That also includes support for VR audio, spatializing, assets (everything from basic objects to more complex systems, materials, environments and terrains) and many of these for free. In this department, Unity is a little bit ahead of Unreal Engine. Unity Asset store contains more plugins and content. Moreover, the communities behind both engines create a perfect support for solving various issues and allow for the most elegant solutions in the design.

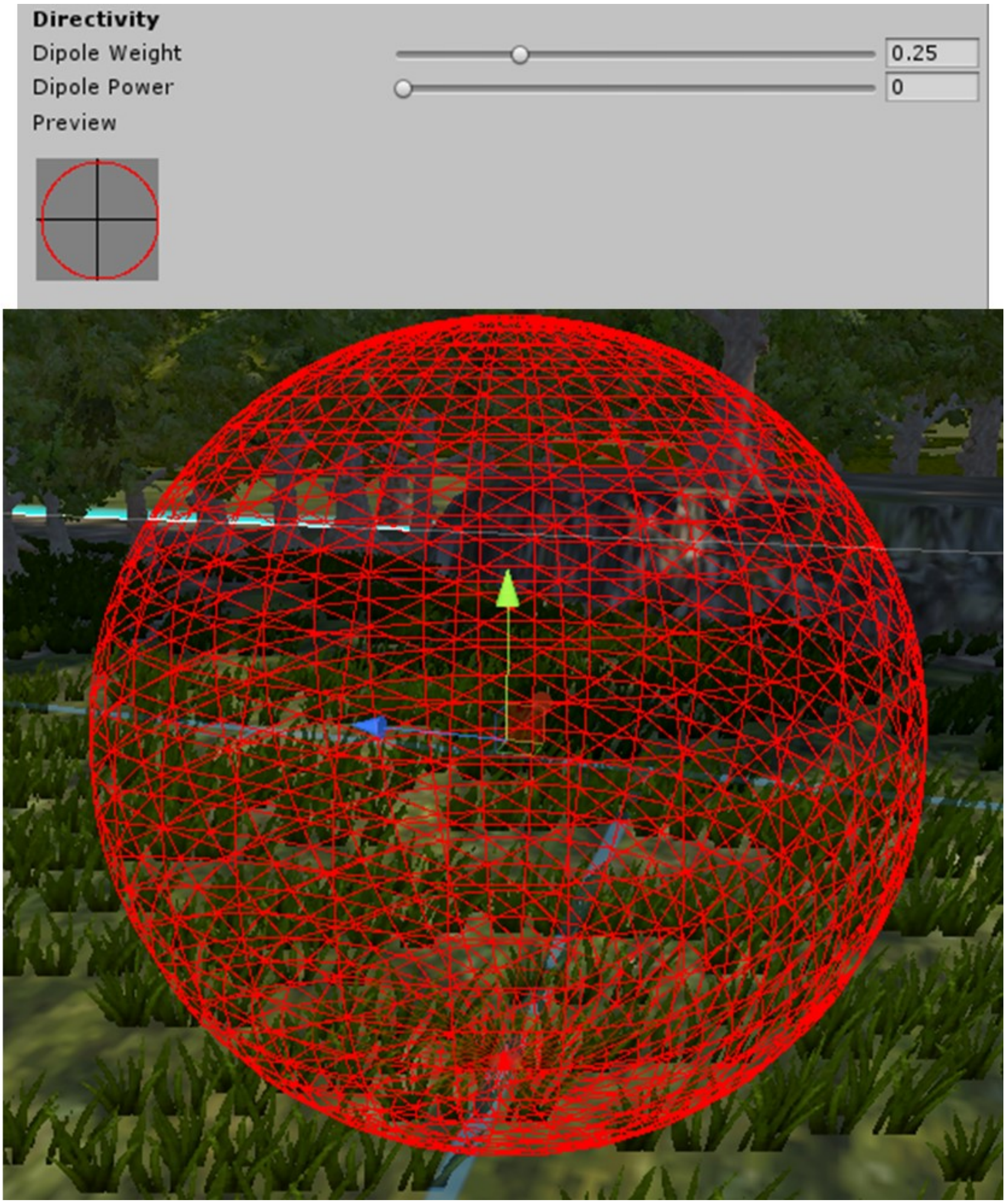
Regarding VR development, there is, again, no definite decree. According to [10], which is in agreement with some other opinions, it all depends on the expected result. Unreal Engine offers more support for beginners as well as deeper details, better graphics, but consumes more time and resources to achieve the promised greatness. Unity is on the other hand more suitable for smaller budgets, smaller projects and less demanding games/applications.

Additionally, one of the key aspects of my application is the possibility of exchanging HRTF banks that are used to generate spatial audio. Both engines possess an audio spatializer that has a generalized HRTF bank, which is used by default. Problem is that finding its parameter specifications is impossible. And in unaltered setting, there is no possibility to switch to a different or a custom HRTF bank. This is luckily solved by Steam Audio Plugin [11] that is freely distributed and downloadable from the asset store.

2.3.2. Steam Audio Plugin

As mentioned above, Steam audio plugin [11] might be one of the most essential parts of the application design. It not only allows specific settings regarding HRTF, but also switching various HRTF banks on runtime. The latter option is basically the most basic specification given by the assignment – the ability to change HRTF sets to test quality of specific/custom HRTFs. There is virtually no limitation on how many different HRTFs can be loaded inside the plugin, which could also allow for immediate comparison of many different banks. Loaded HRTF bank needs to be in spatially oriented format for acoustics (SOFA). SOFA file has been standardized by Audio Engineering Society (AES) and is used for storing spatially oriented acoustic data, which are used for HRTFs, binaural room responses or directional room responses [12].

Besides HRTFs, there are other aspects that can be set up via Steam Audio Plugin. There is a built-in support for different sound wave behavior depending on the material of the surroundings. Due to this possibility, there are various materials like wood, concrete, glass and more, or even a custom one, to be chosen from. Using an object script, there are options to choose the HRTF interpolation – in case a specific function is not defined for the direction the sound is supposed to come from, it can either be selected to choose the nearest one, or to interpolate the four closest functions to calculate the desired direction. A directivity parameter can be set and adjusted. This way, the source can function omnidirectionally or as a dipole. The directivity weight parameter has a range from 0 to 1 and depending on the value, where 0 is omnidirectional and 1 is dipole character, the directivity pattern changes. Directivity power then adjusts the sharpness of the patterns. Examples of different settings can be seen in Figures 2 and 3 below.



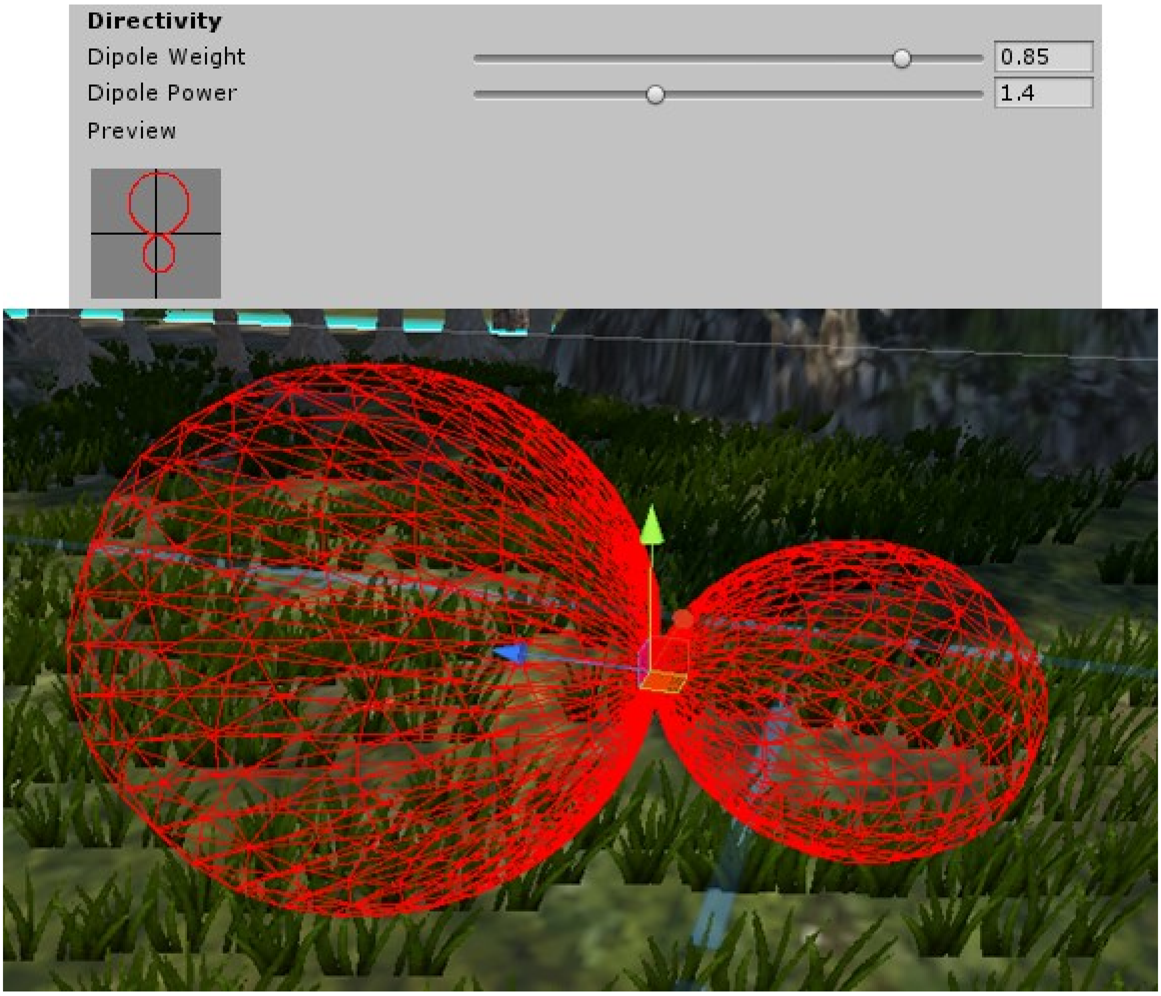
Directivity

Dipole Weight 0.25

Dipole Power 0

Preview

Figure 2: Effects of dipole weight and power on directivity 1



Directivity

Dipole Weight 0.85

Dipole Power 1.4

Preview

Figure 3: Effects of dipole weight and power on directivity 2

An interesting step towards real audio rendering is air absorption that can be turned off and on (no scale for determining the amount of absorption is present) and is dependent on the distance the sound travels as well as on its frequency. Higher frequencies are dampened faster than lower frequencies. And there is plenty of other options to choose from. Occlusions with multiple possibilities of calculations, reflections, physics-based attenuation and more.

2.4. VAS

Virtual acoustic space (resp. virtual auditory space - VAS) [13] is known as a technique, where the sound is presented in a "closed field" (over headphones), while imitating an externalized sound, creating a spatial perception. The illusion is such that we can reproduce a sound originating from any direction in space. In a free-field, e.g. in real life, we localize and perceive sound with the help of audio cues, depending on how sound waves arrive at and interact with our ear. That is also affected by the shape of our head, as the interaural distance and shape of the ear are different for everyone. These audio cues are known as interaural time difference (ITD) and interaural level difference (ILD). This externalized sound is also due to frequency and direction reliant filtering of our external ear structure. In contrast to ITD and ILD, these spectral cues are "calculated" separately for each ear – the same sound source, coming from various directions, will produce different spikes on our eardrum. These spikes and peaks are also very different for every listener and therefore do not form a universal function.

All of these cues combined – ITD, ILD, spectral cues received by measuring impulse responses – make up the so-called head related transfer function (HRTF). This defines the function of filters for sound waves, coming from given direction and allow us to simulate the condition. The array of HRTF impulse responses can be converted into a filter bank of sorts. A reproduced sound can be afterwards convolved with one of these filters and presented over headphones. This will create the illusion of the sound being from an externalized source and is perceived as a spatial sound.

2.4.1. HRTF, ITD, ILD

Head related transfer function (HRTF) is defined as a response an ear would receive from a specific point in space. This response is affected by the constitution and shape of our body and not only what is visible from the outside. Apart from the shape of our head, our ears, earlobes, thorax and the overall robustness and density of our body, the way we perceive sound is also modeled by the inner "architecture" of our body. Parts of the ear, including the tympanic membrane, play a vital role, but also the shape of our nasal cavities as well as the oral cavities. That is also why we usually perceive our voice in a different way than to what others hear – some of these aspects are omitted and others enhanced. HRTF characterizes how would the sound arrive at the outer end of our auditory canal. That is also why one of the methods of measuring HRTFs includes placing a special probe inside the ear in order to receive accurate recordings.

Before going into details of HRTFs, interaural time difference (ITD) and interaural level differences (ILD) need to be defined in detail.

Interaural time difference is an important part of our perception ability. Our hearing is sensitive enough to register time differences between the left and the right ear. It can be imagined as creating a triangle, where the ears and the audio source create the vertexes. Unless the sound is directly in front of us, in front of the center of the head, one of the source-to-ear sides of the triangle will be longer than the other and therefore, the time the sound wave travels along this side will be longer. When the sound reaches one of our ears sooner than the other, it gives us the information about direction the sound is coming from. That is also one of the aspects of binaural hearing in VR. The sound to one ear is delayed behind the other, and therefore creates the illusion of spatiality. These differences are not only due to the location of the sound source, but also because our body is in the way. There is a "shadow" cast by our head and our body, which interferes with the sound wave and delays it.

Along with interaural time difference, goes the interaural level difference, also known as interaural intensity difference (IID). There is a different intensity and frequency registered at both ears, depending on the direction and distance of the sound. The "shadow" of the body and the head plays a key role for this parameter as well, however the interference of our body (mainly head) is more important in this case. Density of our head is also vital. It is worth noting that high frequency sounds are more affected by the shadow our head casts than the low frequency sounds. Now back to HRTFs.

As mentioned above, HRTF consists of interaural time differences and interaural level differences as well as from spectral cues. The time of arrival at the ear and the volume level of the perceived sound is caught at each ear separately and afterwards compared – hence differences. Combination of ITD and ILD, a so-called head related impulse response (HRIR), is used in convolution with a sound in order to receive the spatial representation of the given audio source. HRTF represents a transfer function and is therefore a Fourier transform of HRIR.

We can differentiate between two different HRTF categories, those, where the sound is produced further than 1.0 m from the middle of the head, are called far field HRTFs, and the ones where the source is closer, are denoted near-field HRTFs [14]. Far field HRTFs offer various advantages. According to [15], beyond a few feet, the impulse response (IR) does not change much with distance. For that reason, the same IR can be used for any distance and the sound is afterwards adjusted by roll-offs, absorption filtering and other methods. Near-field HRTF begins where this condition starts to fail. In the far-field HRTF, the center of the space is considered to center of the listener's head. Near-field works with the center of the ear – two ears make it more complicated. Near-field presents different approximations in regards to ITD and ILD. ITD is in principle absent, because the audio source is too close and the time differences for both ears are way less comprehensible. On the other hand, ILD can observed at much higher levels, for the very same reason the ITD is missing. The head shadowing is increased, which leads to different low a high frequency attenuation. This is also nicely portrayed in Figure 4. There is also a so-called dead zone boundary, which is very close to the head, and in which we cannot safely differentiate the position of the sound. An excellent example of the far-field, near-field, dead-zone transitions is our ability to detect mosquitos.

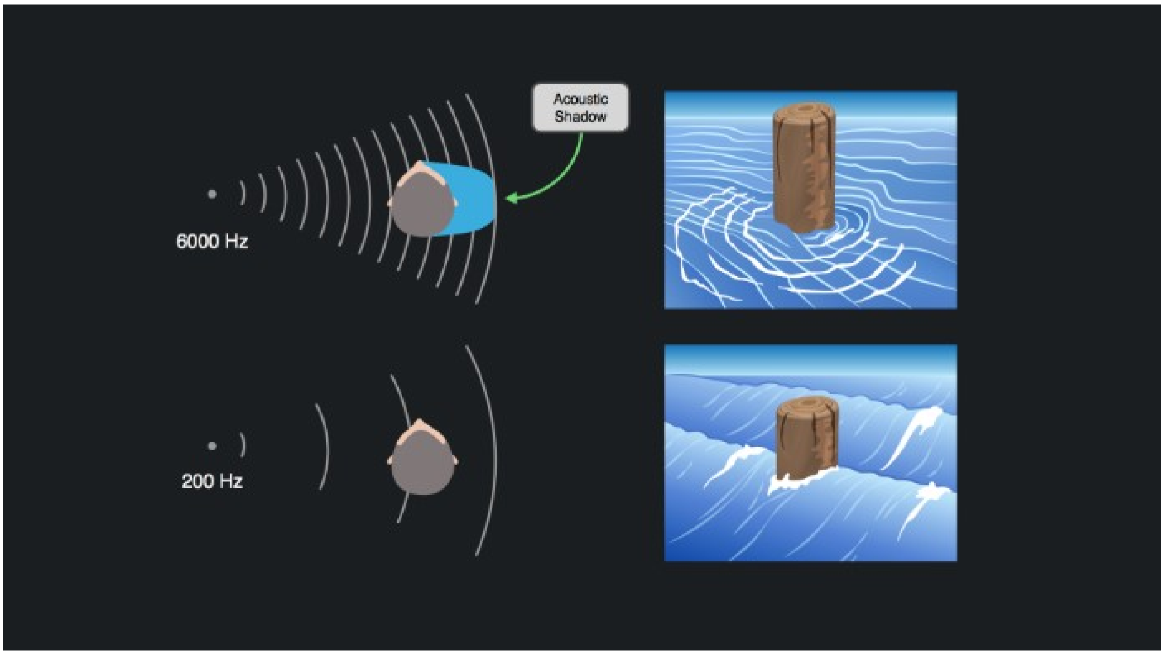


Figure 4: Head shadow effect on low and high frequency sounds [15]

The conclusion is that freely available HRTF banks are field distance specific and therefore, it is important to fit the test specifications to the selected HRTF bank in order to ensure precise sound-HRTF convolution and veracity of the test.

2.4.2. CONE OF CONFUSION

Cone of confusion is a set of points that create an equidistant from left and from the right ear, with the apex originating from the center between one's ears. In this cone, one phenomenon applies and what may occur is the so-called front-back confusion. A reproduced sound that is actually coming from in front of us appears to be behind us or the other way around. That is mainly caused by the ITD parameter, which is the same for either of the front and back sound positions.

According to [16], this confusion often appears in experiments where the source and subject positions are fixed. The same study also proposes and proves that a solution exists. They conducted two experiments, one, where the listener was allowed to move their head and a second one, where the sound source was moved. Using this technique, the front-back ambiguity disappeared. This also proves that it not necessary for the listener to move, but also that the sound source can be move to resolve the confusion. However, it is still necessary to watch out for this in many experiments, where the head and source movements are restricted. Most studies compensate for this error during results processing, by simply excluding answers affected by the front-back confusion.

2.5. AUDIO QUALITY MEASUREMENTS

Apart from specifications of hearing, it is also important to consider all the acoustic parameters that can be measured as a valid part of audio quality tests. There is a wide variety of possibilities and many of them had been previously tested. Virtual reality, however, presents a new challenge and a entire new field for audio quality tests. This section partly sums up the parameters that are being tested and some measurement approaches that are often used.

2.5.1. JND

In full, just noticeable difference (JND). The smallest step of a specific quantity a person can perceive. It can be translated into a smallest angle, to the shortest step in quantization we can see in a picture, the smallest volume step in audio we can recognize and so on. In other words, a change in quantity that is below the threshold of JND is not distinguishable by humans.

One of these parameters that is applicable to this study, is minimal audible angle (MAA). MMA is an angle, in reference to the user, which sets the distance in space, between two audio sources, in order for the user to be able to differentiate between them. It was first described in [17] as *"The smallest angular separation that can be detected between the sources of two successive tone pulses (the minimal audible angle) …".* This measurement is directly dependent on the type of the sound used and the structure of the stimulus, and can be measure in both azimuth and elevation [18].

2.5.2. TIME PARAMETER

Another parameter that is often considered, is time (or reaction time). Usually it is a form of time it takes the user to locate the audio source. Experiments like these put the user inside a virtual world and tell them to find the source, by either finding the audio source in space or by tagging the position the sound is coming from.

A different approach is to measure the time it takes the player to find the direction the sound is coming from. That is usually achieved by monitoring the player's field of view (FOV) and marking the task complete once the sound source is in front of the user [19]. A more precise layout of the audio sources is needed for this case, because a random set of sources could often lead to spawning an audio source that is already in the FOV of the player or right next to it. That also leads to the need of eliminating sources that are basically discovered by mistake, e.g. by a very rapid movement of the head from side to side, and calculating for the cone of confusion errors.

2.5.3. Absolute localization

Another possible angle is absolute localization, where a subject is supposed to pinpoint the exact location of given audio source. That can be done using various methods as described by Pernaux et al. [20]. These would be:

a) Using a special user interface (2D/3D), where the subject can describe the sound location
b) The subject has to localize the sound source and get to the spot as soon as possible
c) Laser pointers or other means of selecting a sound origin from distance

There are various upsides and downsides to all of the above-mentioned methods. A user interface is the simplest to implement, but lacks accuracy as it is most prone to human error. The subject is usually sitting still and there is basically no need for a virtual environment. A sound, or a pulse is played and afterwards, the subject is supposed to tag the location in the special 2D or 3D interface. 2D interfaces are implemented by projecting planes, describing the 3D space around subject's head. The 3D interface depicts the space around the user directly. For example, a virtual (but still only an on screen) head model, with a sphere around it, where the subject points the exact location, he perceives the sound from.

On the other hand, the method of "move to location" doesn't need to compensate for the human error, but clearly introduces a different problem. In case a time parameter is involved, the distance of two consequential audio sources needs to be accounted for. Moreover, a proper listening medium has to be chosen, because the usual cable headphones could be restricting and problematic (unplugged cable, noise generated by friction, …).

The final option is using controllers or other connected objects as laser pointers. A ray is cast from the controller in the direct line and a collision is recorded. That collision is then recognized as the point of origin of the sound. The main disadvantage of this method is that the test either has to be set up with no rotation allowed, but that leads to a problematic source tagging, because the user cannot see, where exactly is he pointing when selecting a spot behind him or on the sides. Or with rotations allowed, but that way the test partly transforms into getting the audio source inside the FOV and afterwards, only detecting the precise location in front of the subject. For that reason, measurements like these should be accompanied by recording another parameter along with the absolute position of the source, in order to get a better specification.

2.5.4. Performance enhancement parameters

One of many uses of VR is also teaching and training. Medical, lingual and many more fields are starting to use VR to speed up learning of new skills, precision and practical application of knowledge gained in theory. There have also been military applications, regarding various sets of skills. With all of these, the main purpose is performance enhancement. Tests are often run with increasingly more difficult sets of tasks and the increase in performance is measured. One example would be [21] [22], where they let the test subject shoot incoming monsters and they measured the time of response and the number of monsters that got to the player before being shot down. The expectation was that the response time of the subject would get shorter with subsequent levels, where the monsters flew faster. In military training, the main parameter could also be the number of enemies that would be terminated in a given time spread, or the precision of shots.

In relation to audio measurement, the main parameter would be the response time of the subject, as faster response time would probably be observed with a more precise audio synthetization and positioning.

2.6. SUBJECTIVE TESTS

Subjective tests can have different forms and different conditions, depending on the preferred outcome. However, the lack of standardization may affect reproducibility or fail to sufficiently reproduce conditions of the final product environment. There is also no reference for the results to be compared with and therefore a specific evaluation method has to be set. University of Salford provides some ideas that should be followed during subjective testing [23].

Subjects themselves might be divided into two groups. Naïve listeners and trained listeners, partly depending on the complexity of the test and on the desired results. Derived from that, a corresponding hearing tests have to be conducted and training sessions provided. For less complex tests, training sessions don't have to be created, but in case of VR, some listeners might have never encountered virtual reality headsets before and that might present an avoidable error. The subject group can also be picked depending on the desired focus group of the test outcome and might have to undergo a special (sensory) assessment in order to ensure that they are fit to participate in the study. With the same thought in mind, testing methods have to be rationally selected.

Many methods have been established over the years and all of them present advantages and disadvantages. Depending on the chosen technique, some rules might have to be drawn. For example, lexicons of evaluating statements, scales, comparison pairs and similar. Going into a bit more detail about the evaluation methods, lexicons might contain specific tags for judging the sound parameters – e.g. soft, sharp, unclear, muffled –, while comparison pairs contain two opposing statements – e.g. yes/no, audible/inaudible. Scales on the other hand offer more flexible evaluation, but might present a hard to process results, as their interpretation depends on fastidiousness of every tester. But even scales offer more variations, which give the subject more, or less, freedom. Scales can be made up of let us say five levels (significantly bad – slightly worse – same – slightly better – significantly better) or allow for a full 0-100 scale choice. Optional is also a use of a reference stimulus to compare the samples to.

The environment, in which the tests are conducted plays an important role regarding results and as mentioned above, should in some cases be reproducible. Having said that, anechoic chamber might be a perfect fit for certain experiments, but does not correspond to real life situations. Moreover, all should be defined prior testing in order to ensure, that the conditions will remain the same for every tested subject as variation may affect the results. The choice of the surroundings can also be influenced by the nature of the audio source, e.g. headphones require different testing circumstances than loudspeakers.

Finally, other test specifications should be picked to ensure a successful outcome. The test should not be too long to avoid errors from loss of focus, the audio tracks should be of sufficient quality and appropriately mixed with applied fade ins and fade outs to eliminate disruptive clicks and cracks. Moreover, the test should be intuitive, otherwise the subject might stray away from the main task and the number of subjects should be high enough to produce statistically significant results.

2.7. HEAD MOUNTED DISPLAYS

In order to be able to work within virtual reality, a display medium is necessary. Especially for virtual reality, the choice determines a lot about the functionality and development possibilities. The selection of the device also determines, which plugins, toolkits and other assets would be at my disposal to work with. Along with what setup would be needed to make the delivery device work – e.g. consoles, base stations, screens and so on. There are three possible categories to choose from.

1. Gaming console powered VR – PlayStation VR
2. Cave systems – using a whole room, every wall, for projecting the virtual world
3. Head mounted displays (HMDs) – an on-head display, powered by either a smartphone or a computer

According to reviews, PlayStation VR actually offers a nice player experience. Also, thanks to the design of the console, which is purely performance based and its sole purpose is gaming and movies, the VR headset is capable of handling advanced graphics and applications without a hitch (which can be a rather demanding task for many computers and most graphic cards). Nevertheless, PlayStation VR does not offer simple support for programmers and many unnecessary difficulties might appear during the development process.

Cave systems might look very impressive, but they need to be backed up by a substantial computational power. Apart from that, they require a lot of room – a room, to be specific – and quite an expensive setup. They are suitable for specific types of experiments and for well-funded applications.

As mentioned above, HMDs can be split into two categories. One that uses smartphones as their source and the other that is connected to a computer. Smartphone-based HMDs, undoubtedly offer less performance, therefore a poorer quality of audio-visual stimuli. However, they are very nondemanding in the matter of setup. Most of today's middle-priced smartphones can power the headset, and apart from a controller, nothing else is needed. The smartphone is inserted inside and everything else is taken care of. On the other hand, head mounted displays offer as much power as the computer that powers them. That enables much more sophisticated graphics and audio reproduction. Also, they are more open to testing, more complex app development, and support various plugins to make the process easier. The downside is the price and mobility in comparison to the latter type and the fact that not all computer graphic cards are capable of running VR applications. And even among these HMDs, there are different configuration demands. Based on the technology of tracking, some additional setup might be needed, e.g. base stations.

Nevertheless, computer powered HMDs present the most viable option. Currently, there are three main branches to choose from. HMDs based in SteamVR, Oculus VR and HMDs based in Microsoft Mixed Reality. All have certain toolkits and libraries to simplify it for beginning programmers, who are not able or do not have the time to write every behavioral pattern and interaction for VR themselves. Moreover, there are downloadable plugins and assets that can improve audio handling or graphics content.

When it comes to downloadable content, toolkits, plugins and assets, Microsoft and Oculus cannot measure up to Steam. After working with two of them, Windows Mixed Reality Toolkit (MRTK) and SteamVR, SteamVR proved to be more complex and dependable. MRTK contains scripts that are not fully and reliably functional and handle certain functions and calculations in an improper and incorrect way. Moreover, as mentioned in chapter *2.3. Choosing a working platform*, SteamVR offers Steam Audio Plugin which's functionalities would save a lot of time and offer new possibilities for testing.

However, there is another aspect to consider. The actual hardware specification of the HMD. As one of the goals is to link the audio test to graphic representation of the environment, the HMD resolution and quality of the display have to be considered. Various models were reviewed via online sources/reviews and the summary can be seen below in Table 1. Some models were also tested directly by me (HTC Vive, Acer, HTC Vive Pro). Apart from resolution, an audio output also has to be taken into account. Some HMDs offer a regular 3,5 mm jack connector, while others have their own, built-in headphones. Those with unpluggable headphones have to be removed from the selection straight away, because the quality of those headphones would never be sufficient to run high quality audio tests and would only introduce an uncompensable error. The technology/method of determining user's position (= tracking – un/marked inside-out/outside-in) is not important for our tests. Even though we need to know the precise position of the user in order to place the audio source correctly, nowadays methods are all dependable enough that there would be no difference between these technologies. And the testing space will be completely cleared out, ergo, there are no disadvantages to either of the possible tracking methods.

Table 1: Overview of viable head mounted displays available on the Czech market; prices are average possible prices, short term sales not included, from reliable suppliers (collected during Fall of 2018), as of May 2019 during rechecking, some models have already been discontinued

HMD	Display	Resolution [px]	Platform	FOV	Frequency	Headphones	Price
HTC Vive	Dual AMOLED 3,6" diagonal	2160 x 1200	SteamVR	110°	90 Hz	3.5 mm jack	16 000 CZK
HTC Vive Pro	Dual AMOLED 3.5" diagonal	2880 x 1600	SteamVR	110°	90 Hz	built-in	36 000 CZK
Oculus Go	LCD 5.5"	2560 × 1440	Oculus	95°	60/72 Hz	3,5 mm jack + built-in loudspeakers	7 000 CZK
Pimax 4K PC VR	8.29MP IPD	3840 x 2160	SteamVR	110°	60 Hz	3,5 mm jack + removable headphones	9 500 CZK
Acer Windows Mixed Reality Headset	2x LCD, 2.89" x 2	2880 x 1440	Microsoft	100°	90 Hz	3.5 mm jack	11 000 CZK
Oculus Rift	2x PenTile OLED	2160 × 1200	Oculus	95°	90 Hz	built-in	13 000 CZK

2.8. Background research on VAS audio quality and HRTF based measurements

In general, there are papers that are purely scientific, meaning they use special interfaces, specific frequencies and sound setups, often do not use visual feedback or focus on new methods of measurements and audio representation. Others use their research for different reasons, learning, new experiences in VR et cetera and conduct the experiments in another form, either for a specific gaming platform/application or simply use a visual feedback and create a feeling of a more specific approach. My aim is to combine visual and audio perception and the main goal of this section is finding, what has already been discovered, which measurements are common, and therefore unnecessary to re-invent, and eventually, which methods and discoveries are considered a dead end and are not worth pursuing.

2.8.1. Localization and user interface methods

According to [24], which in turn relies on the research of V. Larcher [25], there are, in general, three types of possible tests for measuring quality of VAS in regards to localization:

1. Comparison of two audio sources – the subject compares differences between two static audio sources. These tests are time consuming and might not apply well to VR.
2. Visual representation of multiple sound sources is presented to the subject, who in turn chooses the one he/she thinks the most corresponds with the stimulus he/she is presented with.
3. Absolute localization tests – the subject reports (in different ways), where he believes the audio is coming from. The reporting interface can be constructed in various manners – anything from a simple user interface, where the subject tags the direction/area the sound is coming from, up to a 3D go-and-tag task, where the user moves around the virtual space and in order to localize the audio source.

Despite the fact that the original study [25] was written in 2001, and the paper [24] only a year after, other researches derive exactly from those methods, which concludes that even to this day, the summary is fairly accurate. The same paper [24] also discusses various aspects of VR audio tests in general. One of the things worth pointing out is that there might be a side effect (depending on the goal of the measurement) of feedback to the user. Human mind is incredibly adaptable and if we provide, after every audio source location pinpoint, feedback to the subject, a learning phenomenon might appear. Therefore, the mind itself could compensate for a mistake that was present in the test on purpose. That might not always be the aim and needs to be watched out for.

A more thorough division of absolute localization system is drawn in [20] and has already been described in the section *2.5.3. Absolute localization*. The team conducted an experiment, based on reporting systems used in absolute localization tests in order to assess them in terms of efficiency and accuracy. They ordered the methods from the least to the most effective (based on correct answers). The final assessment was as follows from best to worst: pointing in 3D environment -> 3D computer interface -> 2D interface. However, the results have also shown that even upgrading the computer user interface from 2D to 3D improved the accuracy of the outcome.

Apart from computer-based interfaces, there are also technologically simpler ways to note the user input. Tew and Kelly [18] mention methods, where the subject uses oral description to mark the audio source position. Needless to say, this method lacks precision. Different approach is a physical model of the virtual acoustic space, where the subject can mark the perceived audio source location. That can either be a sphere, for easier calculations, or an entire room/other type of environment, depending on the test composition.

However, a contradictory point is also drawn from the same paper [18] as one aspect is impossible to replace in these tests and that is human error. Tew and Kelly [18] suggest that nowadays, the audio synthesis in general is so accurate, and the quality high enough that it is rather easy to recognize the audio source location, but the problem lies in the human to interface interpretation – human error, the imperfection of our hearing capabilities. Also, the lack of ability to properly describe or mark the spot we really want. In real life, we are usually able to localize the source of the sound very accurately (depends on the character of the sound as well as on the sound frequency), but usually, we do not need to localize the source with the accuracy of centimeters and we usually have visuals to aid us. In audio quality measurements, subjects are expected to distinguish the location perfectly and as most of the subject are not audio quality experts and some may have never even been introduced to immersive VR before, it is almost impossible for them to have that ability and precision. For that reason, the choice of the interpretation system and the structure of the system is crucial in order to minimize this problem.

2.8.2. Measurement methods

For example, Kuppanda et al. [19] created a test, where a subject was supposed to move in order to get an unknown object, emitting sound, into their field of vision. This test consisted of two possible setups. Firstly, the user did not have any visual feedback (the video feed was blacked out) and was supposed to get the audio source in the middle of his FOV. The final position was confirmed with a mechanical click. The second setup consisted of audio cues and visual content, with the task of only getting the audio source inside the FOV. As a parameter, time until the button is clicked was measured for every sample. The aim of this test was to evaluate different sonification methods, and therefore the results were not of importance to this study.

A specific measurement for minimal audible angle (MAA, *2.5.1. JND*) was performed by Kelly and Tew [18]. They created a test with a static audio source A and a moveable audio source B. The test goal was to align both sources by moving B, until there was no distinguishable difference between them. The user was free to switch between them, which was implemented using a crossfade between the two sounds. Moreover, only the source that was currently set as active was emitting sound. The test also did not give any feedback, whether the B source was placed correctly or if there was a reversal error (due to the 180° confusion – see *2.4.2. Cone of confusion*). However, this test was not conducted in immersive VR and there was no visual representation apart from a 2D graphical user interface (GUI) on a computer screen. Nevertheless, this study still yielded results of the minimal audible angle between 3°and 8°, depending on sectors, laterality and longitudinality.

In an experiment [26], concurrent to [18], the same group performed a test for multiple audio sources, using the other source as a distractor. Fist part of the paper studied spectral overlapping between two audio sources. The second part used the same experiment structure as described above, while adding a distracting audio source. The task and conditions remained the same and the results were in a form of comparison between both experiments.

A team, with Olli Rummukainen [27] in the lead, remarked that with more degrees of freedom in movement, it is increasingly more difficult to separate audio quality from quality of experience while measuring quality of audio in VR. And in the light of that, it is necessary to create new measuring methods for audio quality in VR. Their test showed, structurally, the most similarities to the test that was discussed for this assignment in the beginning. The user was supposed to locate the sound source, move to the assumed location and confirm the selection via controller. The audio was delivered either through open-design headphones or via loudspeakers that were set around the room. HTC Vive headset was used. To improve the test accuracy, the subject had to put on the headset prior entering the room in order to not be able to visually locate the loudspeakers and to see the boundaries/size of the experiment room. The virtual environment consisted of infinite dessert panorama and the only visual clue the subjects received, was the border of the room area, depicted by a blue cage (a default Steam VR setting), where the sound source was supposed to be. After locating the current sample, a new starting location was shown and another sample was launched afterwards. The aim was to locate the audio source as fast as possible and to mark the spot using HTC controller.

An entirely different study, by Calle and Roginska [28], created a game, where the goal of the subject was to localize a sound source as fast and accurately as possible. The environment was made up of a sphere and the sound could spawn anywhere around the user. They instructed the subjects that accuracy, time and speed are all equally important and they focused on results on following questions: "How long does it take for the subject to find the source? What is the average reaction time after the sound starts playing?

How fast do we rotate our head in order to find the source?" They are however not the first ones to conduct such an experiment. Fang Chen's [29] experiment consisted of localizing 3D sound, using HRTF for rendering, while focusing on accuracy and time. Moreover, he decided to separate results for male and female listeners. When combined, the average reaction time was around 5.5 seconds and the localization time was around 14.7 ± 9.8 seconds. The conclusion of this study mentions that this research is useful for VR content designers as it is a clear indicator of the required sound length, in order for the player to be able to safely locate it. Final results show the ability of trained subjects to locate sound source in 3.7 ± 1.8 seconds, with the average error of 15.4 degrees. But they also recommend that this is not in an average user's capabilities and advise to aim for the sound duration nearing 7 seconds. The average reaction time was 0.2 seconds.

They also state that a binaural synthesis is a number one technique, when it comes to VR sounds rendering, because of its efficiency and low requirements when it comes to equipment and handling. Another question they only bring up though is, whether a personalized HRTF is significantly better than a generalized one, while drawing from a different source, that even with a generalized HRTF, people's performance might improve over time – the ability to learn and adapt to a non-ideal HRTF bank.

2.8.3. HRTFs and sensory integration

The sufficiency of generalized HRTFs is a direct topic of paper which's name starts with "Generic HRTFs May be Good Enough in Virtual Reality" [30] that was partly co-authored by Microsoft research division. They state that our ability to identify sounds in space is mainly determined by three acoustic cues: ITD, ILD and acoustic filtering (shadow of our body, the shape of our ear and head that influence various spectral cues – see *2.4.1. HRTF, ITD, ILD*) [31] [32]. All of these are rendered via HRTF, which can either be individualized or generalized, depending on whether it has been measured on one person specifically or whether the parameters have been generalized and measured using a head model (as for example on some of these HRTF banks [33]). Problem is, creating a personalized HRTF is very time consuming, expensive and is basically only a one-user solution. On the other hand, generalized HRTFs are never a perfect fit and might therefore present a certain error or even a discomfort when presented to some users. Although, they can be pre-calculated, which makes them easy to deploy on various devices and in different applications.

This paper also directly attacks the same issue as my assignment and that is the multisensory integration and multisensory learning. Using various sensations of the human body, the perception of an unperfect audio/video presentation can be improved/influenced [34] [35] [36] [37] [38]. Therefore, by engaging the visuals, the audio perception can be, to some extent, significantly changed – a commonly seen phenomenon that proves this claim is the ventriloquist illusion [39]. Furthermore, it has been found that a so called "ventriloquism after-effect" appears after being repeatedly exposed to the ventriloquist illusion [40]. Meaning is that one's acoustic space perception can be distorted by the illusion and we may afterwards identify sounds from a different location – a real source is virtually not aligned with what we perceive [41] [42]. This is also called a remapping of acoustic space. Individuals apperceive the audio coming slightly to the side of the actual position. This phenomenon only applies to purely acoustic sensation with no visual cue [43]. Similar effects have also been discovered by the same team [43].

When viewing moving objects that are changing their depth, the auditory system is remapped and we may identify spatially static audio sources as moving objects. Along with a different team [44] they directly confirm my assumption that audio and visual stimuli strongly influence one another a therefore a research into AV distractors in immersive virtual reality is not only valid, but very current.

The paper itself [30] aims to examine the possibility of recalibration of the human perception to the environment (= in order to use a generalized HRTF), rather than to adapt the environment to the user (= a necessity for personalized HRTFs). Using a visual stimulus and impact auditory stimulus, associated with physics of moving objects, they tried to remap user's perception. To confirm, they created experiment that would afterwards test the effect on the ability of acoustic localization. The yielded results prove that it is actually possible to induce measurable improvement in spatial localization under specific circumstances. Their conclusion is that using synchronous multisensory stimulation might lead to remapping the brain in a way that a generalized HRTF would be sufficient and would therefore omit the need to measure individualized HRTFs. They also suggest that a certain knowledge of the sound is also important, as we tend to hear a specific sound of interest even in a noisier environment – also known as "cocktail party effect" [45].

A similar experiment, engaging in significance of visual content's effect on perceived audio quality was conducted by Rummukainen et al. [46]. The premise of mutual influence among seeing, hearing and movement is built on the fact that our experience in the real world is connecting all of these data together, constantly, throughout everything we do. This mental reference is therefore what we should strive for in VR [47]. In immersive VR, it is therefore impossible to separate audio quality measurement from the multisensory experience, without losing some of the test's value [48] [49]. An obstacle VR tests are faced with, is the reference item. Usually, it is an ideal state, but in the case of immersive VR, the reference is the real world. For that reasons, HRTFs play a vital role in audio quality tests.

The paper itself has the goal to study the effect of visual content on the audio quality in VR and illustrate the importance of visual content in audio focused testing in VR, in a 6 degrees of freedom (DoF) environment. Two scenes are tested, one is an indoor setting of a room, with a loudspeaker as a visual cue and the second is an outdoor beach scenery, with the possibility of a moving object inside the scene. Their tool of choice is Unity. Volume roll-offs and distance factors were properly setup to ensure credibility of the environment. Also, different renderers were used and tested in the matter of user's ability to discriminate between them (not to assess their quality). Participants were to evaluate the overall quality of the audio reproduction with respect to the visuals and their movement. The scale was set to range 1-100, and the goal was to induce a feeling of casual and believable world.

Subjects were split into a naïve and an expert group to compare the differences in interaction and perception of the scene. As expected, expert group was more critical. Moreover, they were able to distinguish between different renderers, which led to bigger distortion of the visuals in the scene. This can be taken as a direct proof that these tests are strongly influenced by the test group as well as by who is going to be the intended focus group for the result application. The outdoor scene with the moving object received the highest score, which unintentionally (not the aim or interest of this paper) proves that distractors and increased attention requirements play a role in our perception.

In conclusion, the overall audio quality is widely determined and affected by the visual content of the environment. Despite the direct instruction to only evaluate the audio quality, results show that renderers and the scene content affected the overall quality score. This finding confirms that AV integration is a current field of interest to VR development and cognitive sciences and as reported in [50] [51], it offers significant benefits to object detection, localization and spatial awareness. On final note, they state that a reference free audio quality tests are a viable option for immersive VR. Even though the results might be harder to analyze, it offers various advantages to regular reference testing. Firstly, the focus on one item at a time increases the immersion effect of VR, which could be easily lost if various items were switched constantly. Secondly, creating a meaningful reference for immersive VR is rather difficult and subjective. Thirdly, creating a reference could distort the rating scale, because it sets the maximum possible quality beforehand. And lastly, a specific reference can bias the subjects to traits and features that would otherwise be of no interest to the user. Especially in a multimodal (AV connecting) scenes.

In the end of [27], it is noted, taking into account a different source [52] that HTC Vive may not be ideal for certain measurements. This study [52] is specifically dedicated to test the research capabilities and precision of HTC Vive headset. Vive's latency is very low (22 ms) and its tracking capabilities are very fast and accurate. However, two major problems were discovered. Firstly, HTC Vive uses a reference plane that can be tilted in comparison to the ground plane. Secondly, this tilt changes every time the headset loses its location/tracking and has to find the reference ground again. That may lead to incorrect measurements of roll and pitch and changes in eye-height. The tilted reference ground could even sometimes be observed by regular users, where their playing field appeared slanted. This can suggest that high-precision tests can be affected and a correcting method, or at least an increased caution, might have to be implemented. The downside is that correcting at the beginning of each test would only require a correction method, but the tracking can be lost at any time – the user can accidentally walk out of the tracked area or cover the sensor with his hand while using controllers – and these cases are hard to detect and present a challenge. Depending on the final test proposal, these discoveries may have to be taken into consideration in the end design of my experiment. On the other hand, the main question is, whether these errors would even pose a valid problem in my case. Niehorster, Li and Lappe [52] state that problematic test fields would be human locomotion, travel distance tests, optic flow experiments and similar, which are not part of my application.

2.8.4. Distractor implementation

Olk, Dinu, Zielinski and Kopper [53] published a study, pursuing the impact of visual distractors inside the user's FOV as well as in peripheral vision. Their task was a visual search in immersive virtual reality. They used the knowledge of human psychology and focused on the effect of distractors on our attention and abilities while locating items in space. Their premise is built on the fact that our attention resources are limited, which can be easily seen in our everyday lives. Distraction in general is of interest to many scientific fields and to industry in general. Understanding how easily, why, when and under which circumstances we get distracted may be fundamental not only for many safety features, but also for improving our efficiency in everyday tasks.

The aim of this paper was to create a test of type "find a target among distractors". Most often, tests like these are conducted on paper/computer, in a form of searching for a letter among other letters. Let us say, searching for S among X's. The important aspect is, whether the target shares certain traits with the distractor or not. For example, it would always be easier to search for letter K among Os, than among Ns, as the target letter would be more difficult to distinguish. To increase the difficulty of the task, the search array may be flanked by a different set of distractors that might also either share features with the target or be very different. Two terms are connected with this type of test – congruent, meaning the flanking element is the same as the target element, and incongruent – basically the opposite. The flanker is in a similar category as the target, but the opposite element. As I mentioned before, those tests are usually done on computer screen or in printed version. They are easily controlled and can be systematically manipulated in order to receive a specific result. However, these forms and tests are irrelevant to immersive VR and hold little to no ecological validity. For that reason, Olk et al. wanted to research the distraction in an environment and using tasks, which are close to those in our lives. Ideally, users would be tested in everyday situations, and even though it is not really possible, immersive VR provides the best tool possible – 3D immersive environment, freedom to create any scenario one might need and a certain haptic feedback from the controllers.

This paper directly confirms that some studies have already been conducted in the field of attention research using VR, but these are rare and do not usually focus on ecological validity of the research, target-distractor discriminability and response competition. Aiming for the best contribution, Olk et al. used daily objects as stimuli in their test. Claiming that this study is one of the first of its kind, they decided to follow previously conducted tests and use a similar test composition while improving the ecological validity. The task consisted of a circular array of 6 items on a kitchen table. The subject was supposed to pick the desired item as fast as possible – reaction time was measured. A 2D version of the test was compared to the 3D VR test and various version were created. The first experiment was only the aforementioned circular array, where only one item was to be selected and other items served as distractors. The second test consisted of the same array, but next to the array (sides/top/bottom) was a flanker item that was either congruent or incongruent – can be seen in Figure 5 below.

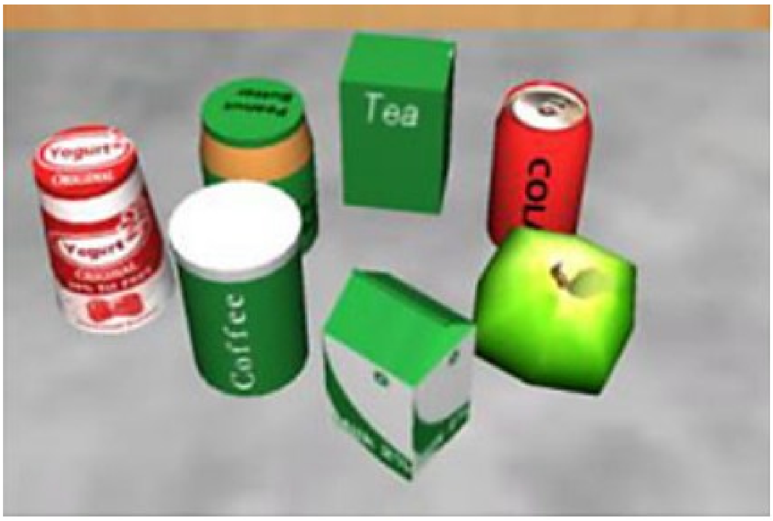


Figure 5: Example of circular array setup [53]

To further improve the experiment's worth, eye tracking feature was added. They measured, how often would the user look at the item before selecting it. But moreover, how often would the subject also notice and look at the flanker – the distracting item that was not part of the circular array. Details like depth were considered, so the size of the items was adjusted to their positions. Their results were processed for different test configurations, for congruency and incongruency of the flanker and yielded reaction times and eye tracking data. Based on their results and several predictions it was possible to observe how the flanker affected the subject's attention depending on the type, but also depending on its position. Interestingly, as the items were daily objects, a specific curiosity was spotted. When the subjects were very familiar with the object, they tended to ignore the flanker more than when the item was unfamiliar, e.g. when a musician would have a guitar as a flanker, his attention would be less affected than by a syringe. Moreover, the distance of the flanker had some effect as well (depending on the position, top - far, sides - intermediate, or bottom - near, with regards to the array). In conclusion, they suggested that future studies should be more focused on real world approach than on scientific tests that lack ecological value. Finally, on an important note, their test was conducted using CAVE VR system and not regular HMD, which presents a very different challenge to the user than an immersive VR presented by head mounted displays.

2.8.5. *Background research conclusion*

Many studies were conducted on the topic of audio in virtual reality. Despite that, not many studies actually connect visual and audio stimuli and even less use real-life audio samples. However, some studies agree on the fact that there is a lot of potential in this field and it is worth pursuing. It has been proved that there is a strong connection between audio and visual stimuli when it comes to multisensory integration in VR. Also, the viability of distractors had been confirmed and some studies attempted to prove its ecological validity and suggested that future studies should be conducted. Although, the topic of impact of visual distractors on audio has yet not been thoroughly researched, it is now becoming more and more relevant as the extent of virtual reality expands.

3. SUBJECTIVE TEST DESIGN

3.1. INITIAL DECISIONS

Prior creating the first test proposal, decisions had to be drawn from the research in *2. Theoretical introduction*.

3.1.1. HEAD MOUNTED DISPLAY SELECTION

Based on the initial comparison in *2.7. Head mounted displays*, a headset was chosen depending on the price, availability in our country and other respective parameters. There are basically only a few models in the affordable price range that are regularly sold in the Czech Republic. Most of the parameters do not differ too much among models, so one of the most important ones is the supported development tool. When it comes to the "battle" between SteamVR, Oculus VR and Windows Mixed Reality Toolkit, SteamVR is definitely more user friendly and better supported platform, which leaves out all of the MRTK and Oculus VR headsets (e.g. Acer HMD and Oculus Rift). Additionally, all headsets that have built-in headphones were automatically out of the question, as their quality is not sufficient for testing. Tracking method is not important, as I am not limited by space for testing and the experiment is not going to be anyhow dependent on it.

In conclusion, HTC Vive was chosen as the best fitting device and its parameters are summed up in the Table 2 below [54] [55]:

Table 2: Specifications of HTC Vive headset

Screen	Dual AMOLED 3,6" diagonal
Resolution	1080 x 1200 px per eye (2160 x 2000 px)
Refresh rate	90 Hz
Field of view	110 deg.
Sensors	G-sensor, gyroscope, proximity, lighthouse type stations
Tracking	SteamVR tracking, inside-out
Connections	HDMI, USB 2.0, stereo 3.5 mm headphone jack
Input	Integrated microphone
Eye relief	Interpupillary distance and lens distance adjustment

3.1.2. PROGRAMMING PLATFORM – UNREAL ENGINE VS. UNITY

As indicated in *2.3. Choosing a working platform*, the final decision came down to Unity versus Unreal Engine. Both are undeniably unbelievable working tools that would serve its purpose, hence the main difference is the programming language they are powered by. As one of the prime parameters for this assignment is time and efficiency and I have previously worked with Unity in a different class and in my free time, the conclusion is that Unity is a more suitable solution. There is a complete support for SteamVR, free assets to improve my graphic content, Steam Audio Plugin and HTC Vive is therefore fully compatible.

3.1.3. Initial test vision

Primarily, it was necessary to decide, which parameter(s) would be measured, emanating from the background research. Secondly, the user input interface had to be selected and implemented. Thirdly, an environment along with suitable audio tracks had to be devised in order to create a believable world. The last step is to try and asses the initial test and draw conclusions for the final test version.

The goal of this study is combination of audio and visual stimuli in order to see, how they affect the results of given measurement. The selected parameter will shape the rest of the test and the environment and as the aim is to create something that will not simply be a copy of a different study, selection of this test attribute might be the most important one. As discussed, prior the assignment, the initial vision was to focus on spatial localization, possibly in connection with HRTF. Reaction time parameter can be measured as a biproduct and there is no need to aim for it. The "go and tag" approach was declined right in the beginning, because it would require a lot of space or implementation of teleportation, which could disrupt the continuity of the test and affect the results. Therefore, a different version had to be devised in order to test the ability to locate the source in order to find out how accurate the sound position rendering with given HRTF is, along with a possible error that people are not able to safely detect. Further specifications will be decided after the initial stage.

In *2.8.1. Localization and user interface methods* there was an overview of possible user input interfaces and their ranking in terms of accuracy and efficiency. Even though a regular 3D environment would be possible in our case, it could present an unnecessary error and in order to achieve clarity of results, a fully virtual user interface was picked to be the best solution. The "gun in hand" option seemed to be the best possibility as it also allows precise implementation. However, output needs to be logged for additional processing, but the execution was not set beforehand.

Next, the environment needs to be believable in order to achieve immersion (to the maximal extent the short test will allow). Key parts of this aspect, in my case, will be simplicity and familiarity with the surroundings. Concurrently, if the audio sources have a visual form in the next stage, the environment will need to be prepared. Because of that, the selection is thinned, as not all 3D object models are easily obtainable (modelling is not part of this assignment and would take up a lot of time).

In summary, first vision of the test consists of Unity environment, implementing Steam Audio Plugin (with the aim to assess its possibilities). The test will allow for logging the test subject's name and answers, along with whether they were correct or not, in some form. The environment will be a calm open-air scenery, including trees, sun and some terrain. The goal is to implement the "gun-in-hand" mechanism, with the possibility to mark a certain location as the answer and adding randomly playing audio sources and locating them via HTC controller and the aforementioned method. Implementation of teleportation will be included in order to see, whether it would have any merit for this task. The test should have a working experiment flow and allow its full completion.

The aim of this first test design is to create a first VR testing application, confirm our ideas and asses the overall feeling. The next stage after that will follow up on the conclusions made from this said first design and implement more complex test structure along with selected measurements and distractors.

3.2. First test design

This section should introduce the first test design and its evaluation, conclusion derived from that and outline changes that will be implemented in the next stage. This entire test is designed, debugged and programmed directly by me.

3.2.1. Environment layout

One of the goals of the first test design was to create a believable and pleasant environment, judge the feeling of the scene in immersive VR and try the controllers in terms of how they react in VR. And to assess the overall presence in the virtual world, the interactions and possibilities. Also, what is and is not possible to cover and measure in the given time.

The entire application was created as a 3D Unity project. The project was setup to support virtual reality technology. In order to get the Steam virtual reality support with their basic assets, *SteamVR* [56] had to be download from the Unity Asset store. Along with that the *Steam Audio Plugin* [11], which supports interchanging HRTFs, Steam base audio spatializer and other audio source settings. As mentioned in *2.3. Choosing a working platform*, there is a lot of assets available in the Asset store. Including this, there are specific packages that can simplify and speed up creation of environments. One of these is the *Standard Assets* [57] package from Unity Technologies itself.

The scene presents natural surroundings, where the main playing area displays a meadow, surrounded by trees and heightened terrain to clearly confine the user (Figure 9). To improve the immersion, much more extensive terrain was added, surrounding the focused player space. This ensures that there will not be an empty “gray area” visible through the tree line (Figure 10). To add more credibility, some grass was spawned, more than just one type of tree was used and the terrain was diversificated. The player is spawned in the middle of the meadow and has a full freedom regarding movement within the real-life playing area and can turn to any direction. A very limited possibility to teleport (taken from the *SteamVR* assets package) is included as well, with the goal to assess, whether it should be present in the next stage or not.

FIgures 6, 7, 8, 9 and 10 below depict the overall timeline from creation of the project up until to the end point of environment design. The inspector (left side of every picture) illustrates the progress of scene content.

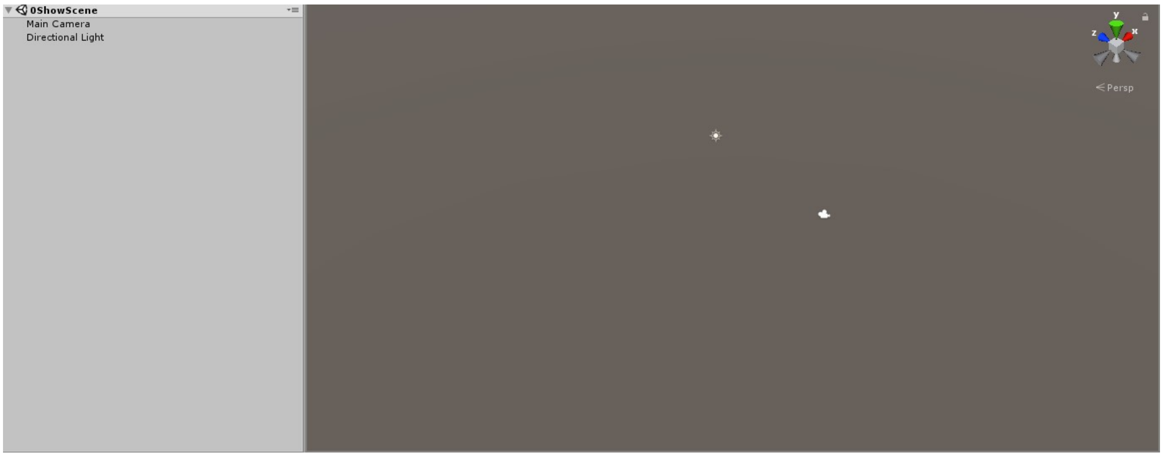


Figure 6: Initial state of the scene

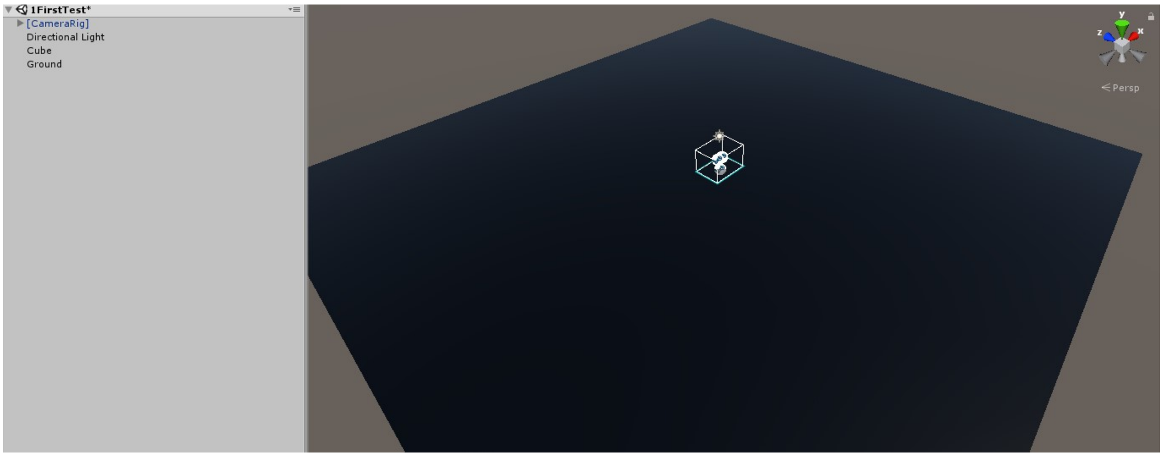


Figure 7: Initial VR rig

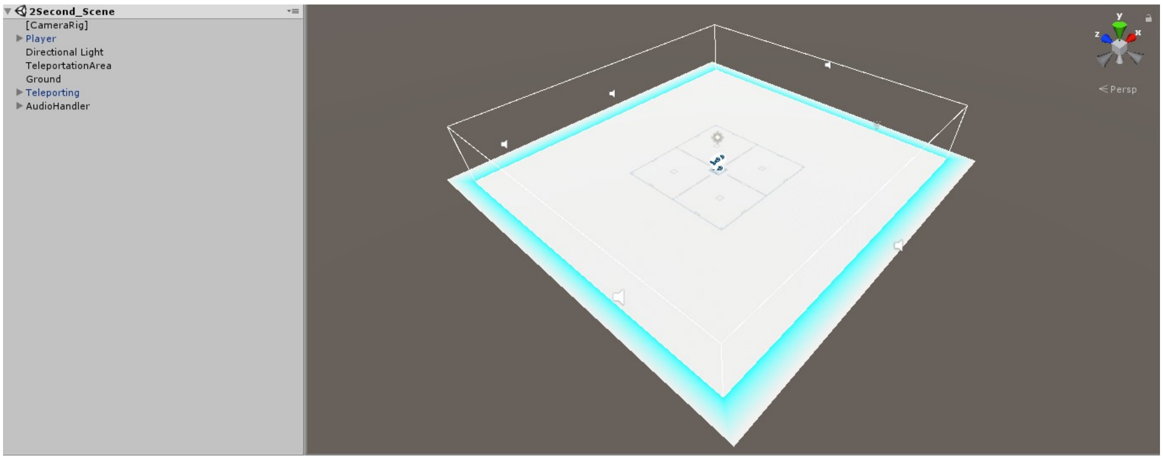


Figure 8: First fully functional VR scene + teleportation

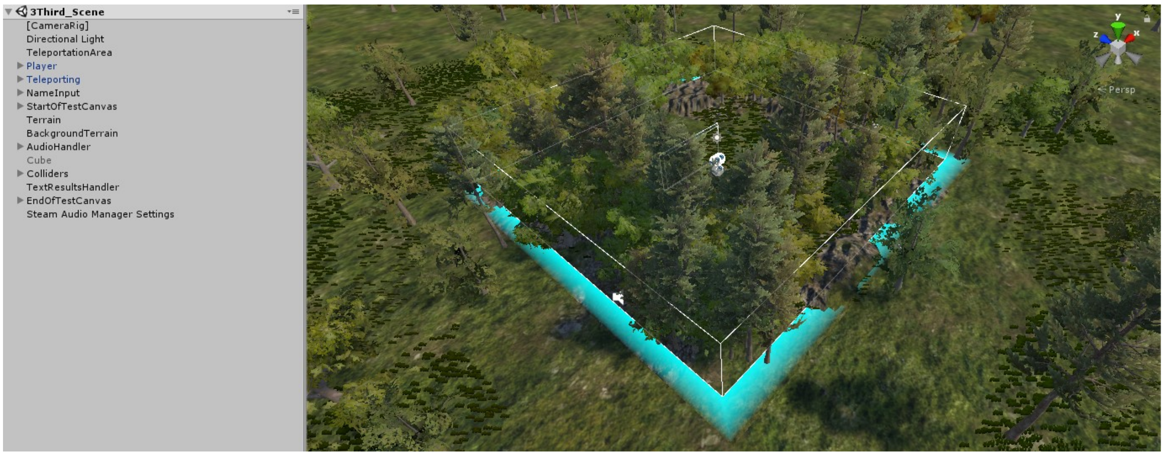


Figure 9: Player area full setup

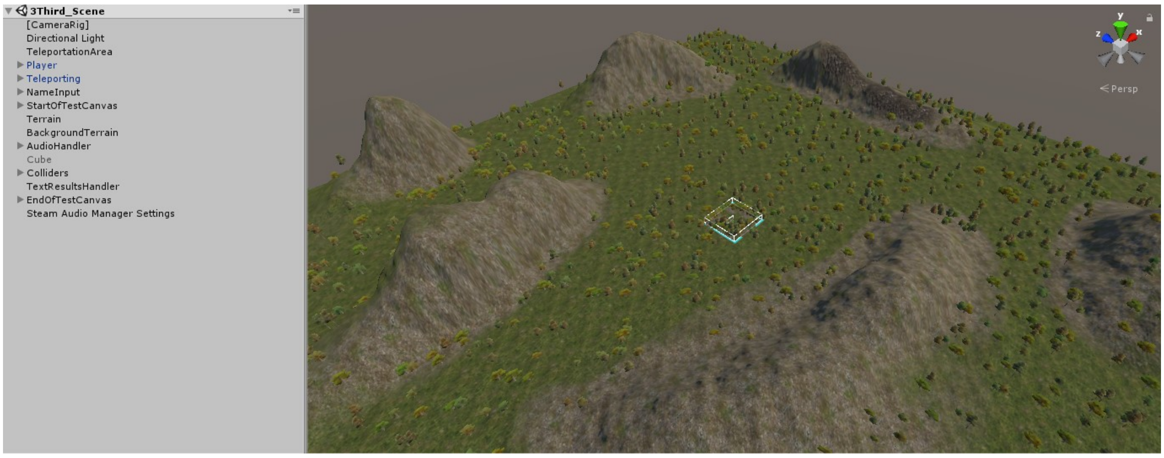
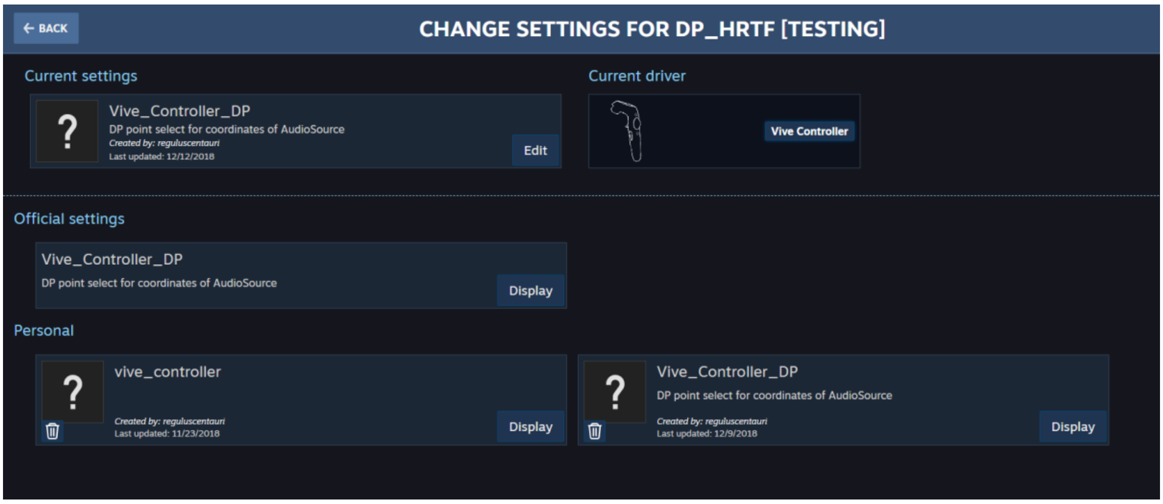


Figure 10: Full scene setup including the surrounding terrain

3.2.2. LOCATION SELECTION AND CONTROLLER SETTINGS

An important first aspect was changing controller input settings. Every controller button is mapped to a specific basic function, but there is also the possibility to assign a custom functionality with a specific trigger (hold, double click, triple click, …), which allows for a wide spread of interactions through controllers. Figure 11 below shows the basic menu with an overview of different setting templates and Figure 12 displays the (left) controller alone. The right controller is located symmetrically on the other side of the screen and can be set up identically or individually, depending on preferences and needs. In order to achieve the possibility to mark the target using a specific controller button, the function had to be created and then manually assigned to the controller. In this case it is a single push of the trigger button, set on both controllers.



← BACK

CHANGE SETTINGS FOR DP_HRTF [TESTING]

Current settings

?
Vive_Controller_DP
DP point select for coordinates of AudioSource
Created by: reguluscentauri
Last updated: 12/12/2018
Edit

Current driver

Vive Controller

Official settings

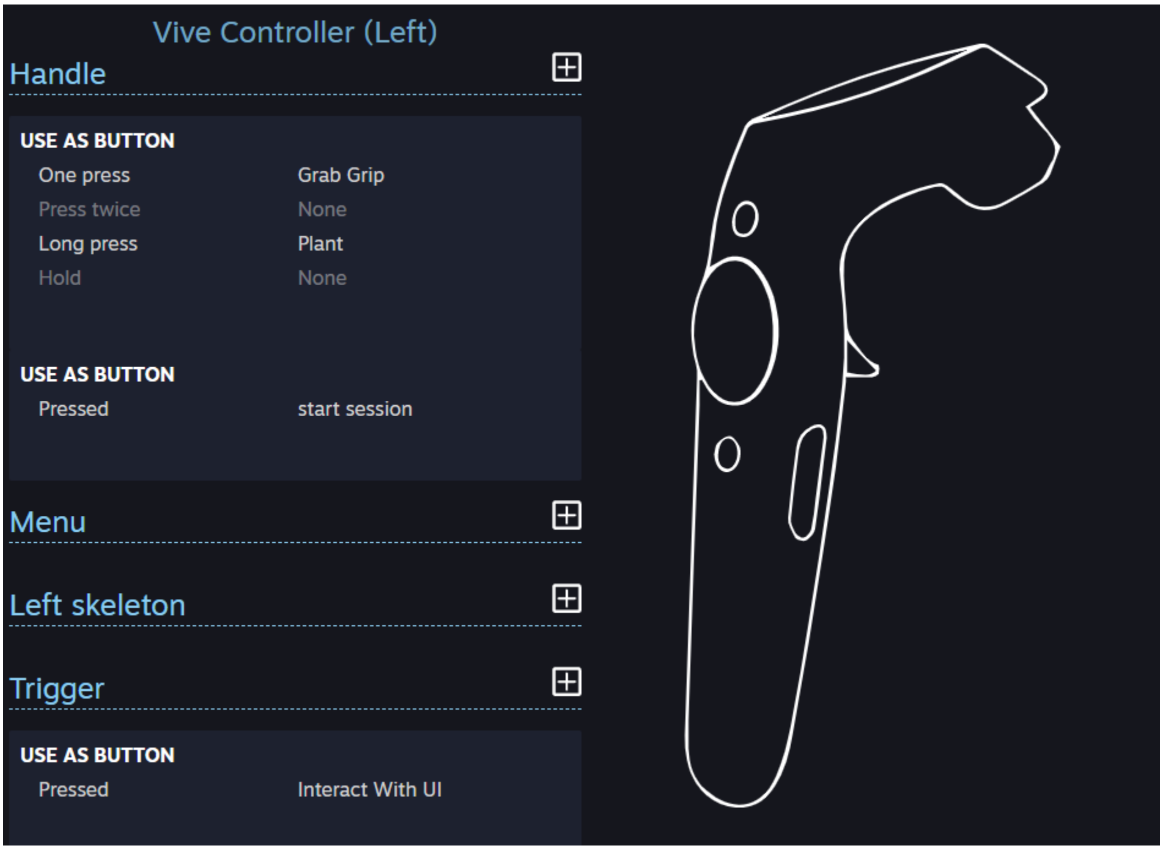
Vive_Controller_DP
DP point select for coordinates of AudioSource
Display

Personal

?
vive_controller
Created by: reguluscentauri
Last updated: 11/23/2018
Display

?
Vive_Controller_DP
DP point select for coordinates of AudioSource
Created by: reguluscentauri
Last updated: 12/9/2018
Display

Figure 11: Steam controller settings menu



Vive Controller (Left)

Handle ⊞

USE AS BUTTON

One press	Grab Grip
Press twice	None
Long press	Plant
Hold	None

USE AS BUTTON

Pressed	start session

Menu ⊞

Left skeleton ⊞

Trigger ⊞

USE AS BUTTON

Pressed	Interact With UI

Figure 12: Controller button mapping menu

In order to log the user input selection of the presumed audio source, the “gun in hand” method had to be implemented. This was done using a raycast that goes in the direct line from the controller. To increase precision, the raycast was accompanied by a laser pointer, so the user sees exactly where he points. Even though only one controller is needed for the test, it is still possible to connect them both. Because of that, different colored lasers were chosen for each controller to avoid mistakes.

Next step was writing the logging function itself. Raycast can interact with objects that have colliders on them and it is the easiest and safest way to record the selected location.

A grid of colliders was constructed around the user area to detect collisions of the rays sent from the controllers. A mechanical trigger on the controller then served as an action button to mark the perceived audio source. Figure 13 shows the visual realization.

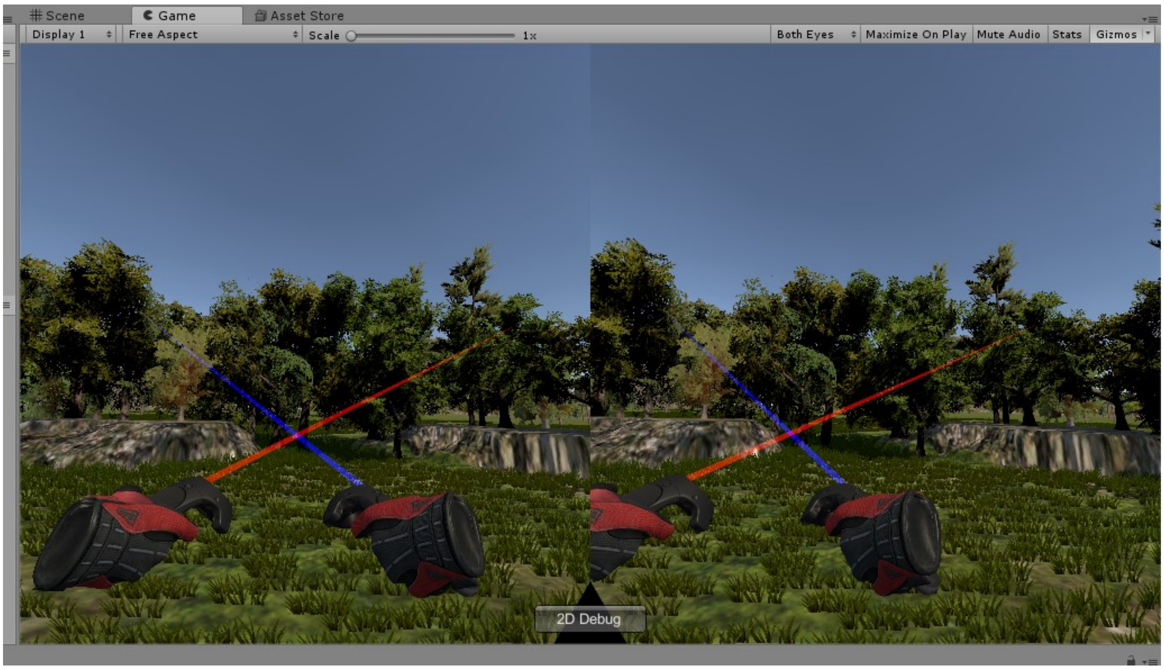


Figure 13: Laser pointer implementation

3.2.3. TEST ARCHITECTURE

The design also needs to feature a first test architecture. Layout of audio sources and an audio managing mechanism that would automate the test, along with the structure of marking the user selected perceived locations.

The audio sources were evenly distributed around the marked area – at the moment a rectangle – that was not visible to the player directly, but was marked by the more tightly grouped trees. An audio manager was created with the purpose of handling the audio sources. For every test run, all of the sound sources were played in a random order, until there were no sources left. In order to make the calculations for marked spot of the presumed sources easier, all of them were placed directly at the border of the area as seen in Figure 14. Only one source was placed inside the rectangle and it was also the only source that was visible (a sphere with a particle effect on it). This source was supposed to serve as a control object – even though the sounds were played in a random order, their locations were fixed and the sounds did not change in between test runs.

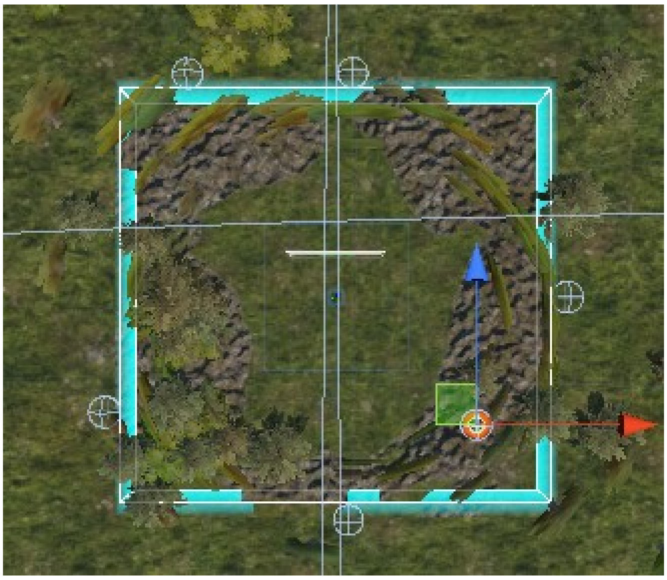


Figure 14: Setup of audio sources

Figure 15 shows the initial audio source setup. Blue spheres mark the distance up to which the sound can be heard. Roll-offs and their character was set up later on, along with spatial character, volume, spatial blend, spread and doppler level. All of these are visible in Figure 16.

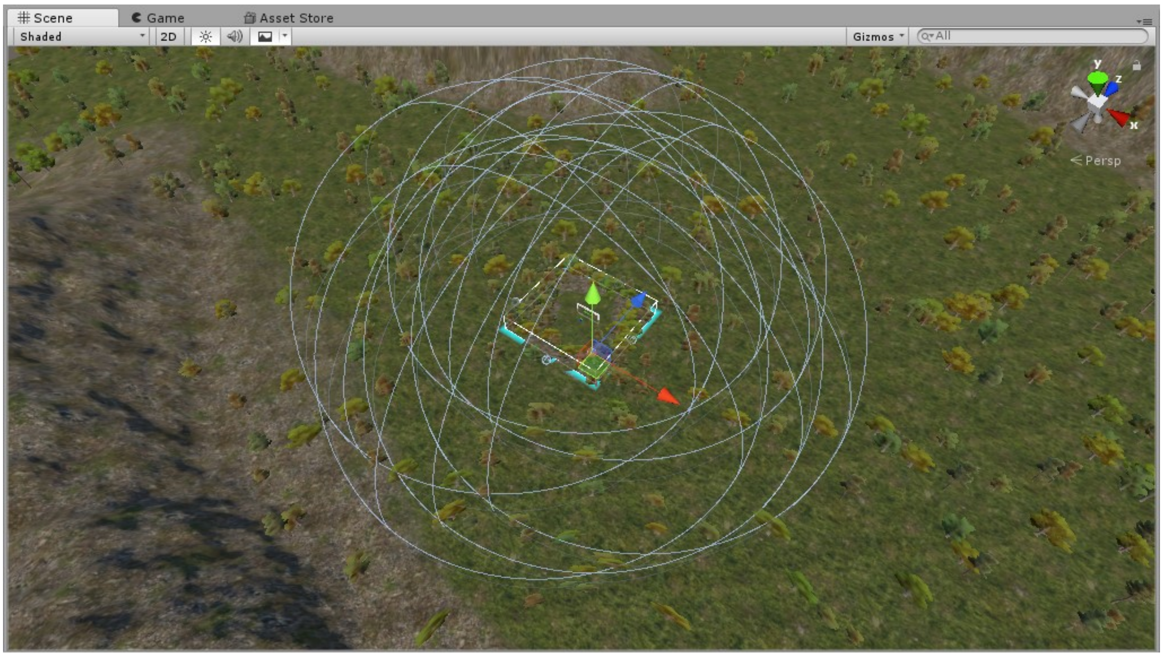
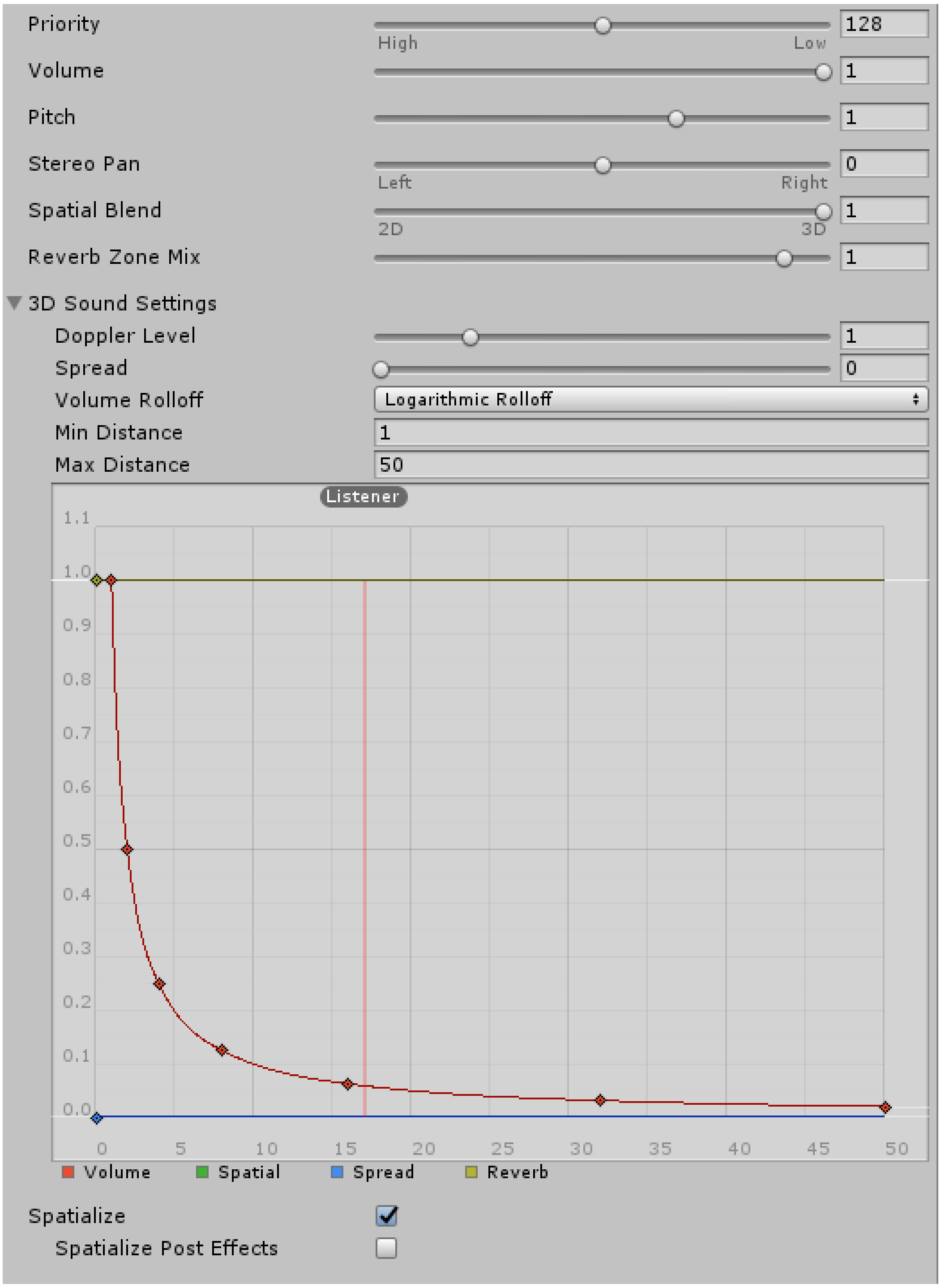


Figure 15: Audio source sound extent



Priority 128

Volume 1

Pitch 1

Stereo Pan 0

Spatial Blend 1

Reverb Zone Mix 1

3D Sound Settings

Doppler Level 1

Spread 0

Volume Rolloff Logarithmic Rolloff

Min Distance 1

Max Distance 50

Spatialize ☑

Spatialize Post Effects □

Figure 16: Audio source configuration example

For audio sources, different music tracks were picked for every source, mainly to allow repeatability during testing by one user (me, as a developer) and also to determine, whether a song is a possible content for the audio test.

3.2.4. TEST FLOW

The subject is spawned in the middle of the area in Figure 9. Firstly, the overseer needs to input the subject's name and afterwards, the testee is asked by a prompted text to push a button to start the test (Figure 17). At the moment, there is no training session and all commands are given by the overseeing person beforehand.

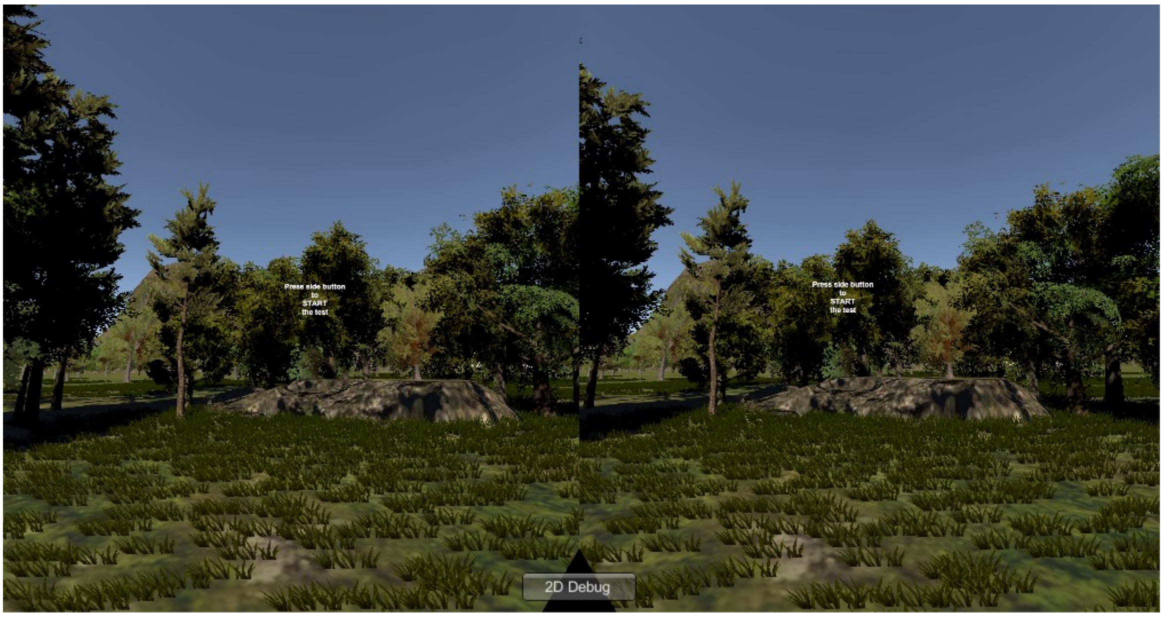


Figure 17: Initial test screen with prompted start text

After starting the test, one of the audio sources is spawned and its coordinates are saved in a remote text file. The audio source keeps repeating until the user marks the perceived location via either of the controllers and this location is logged in a pair with the audio source in a remote text file. Afterwards a fade out is applied and a new source starts playing. This process repeats as long as there are unplayed sources and at the end a new text informs the subject that the test is over. The test then has to be manually shut down by the supervising person.

Multiple tests can be launched in a succession as the text file separates the individual cases. However, the test always has to be manually relaunched, because as of now, it does not allow for restarting after the test is finished. Based on the marked coordinates, number of calculations can be made – position of the player is also known – [0,0,0]. Even though here, a problem can be observed, in case the player teleports himself before marking the location. There is no mechanism that would compensate for the movement of the player or log his location.

3.2.5. *HARDWARE*

The test was conducted using common middle quality headphones, as the main purpose was to test the environment altogether. A better-quality headphones will be used in the upcoming tests. The application was run on two desktop computers:

1. Win 10, Intel i5 8000 series, Nvidia GTX 1070 Ti graphics card, 16 GB RAM
2. Win 10, Intel i5 7000 series, Nvidia GTX 1060 graphics card, 16 GB RAM

Both of these computers are powerful enough to run this application and are "VR ready". There were no problems while running the application on either of the machines and were therefore used for testing and development in the next development phase.

3.2.6. *DISCUSSION AND CONCLUSION FOR THE FIRST TEST DESIGN*

Three people, two audio experts (educators) and I, tested the set up in order to decide, whether the test architecture like this is viable or not. It was determined that at this moment, this test would basically only consist of “get the source in the field of view” and “search the location of the source in your field of view”, which might be usable, but not fully sufficient for the study we wanted to achieve. Therefore, a different approach will be used in the next test design, along with implementation of visual distractors.

Regarding the environment, despite the fact that the trees and the grass are rendered as billboards (their image always turns towards the player), it does not feel distracting in any way – on full details, lower details would distort the overall illusion. That partly confirms that a credible immersion has been met, because the scene does not feel unnatural and also that our sensory perception is limited when focused on specific detail/task, because the testees did not even notice the billboard “phenomenon” until it was pointed out.

Also, a song as an audio source is not suitable for this test, because even with the spatializer, it makes the detection unnecessarily harder and disrupts the main purpose of the entire test. A more graphic content related audio will be used, to ensure the connection between the visual representation and the listening task. However, the audio still has to maintain its content value (no impulse sounds or constant-frequency tone will be used, as these are not usually heard in the real world).

Moreover, HRTF testing possibility will be implemented as well. Current test was using a generalized Unity built-in HRTF set and the following test should provide a possibility to switch for a different SOFA HRTF bank in order to compare the differences for individual test subjects.

As a first test proposal, this structure was sufficient and informative, but a different architecture will have to be put in place for the final test. We were able to evaluate and compare the environment, in terms of understandability, expectations and controllability. However, a more direct approach will have to be taken with measured parameters and HRTF testing, as both of these were put aside during this test design. Moreover, Unity proved to be a fitting tool for this experiment and HTC Vive has not manifested any issues that might introduce an additional error to the test (although a clear testing area needs to be provided in order to secure safe a continuous HMD tracking).

3.3. FINAL TEST DESIGN

This stage will be a final test design for the extent of this work and should provide a reference to whether the application is usable, which changes might be applied in future testing and yield experimental data from tested users. Same as the initial test design, all parts of the application are programmed, debugged and tested by me and those, which are not, are properly credited.

3.3.1. USED ASSETS

This short section will introduce assets that are integrated in the final test design and will give reasons for their implementation.

In order to prove the possibility of testing various HRTFs and changing them on runtime, different free HRTF sets [58] were downloaded. That concludes in three different

HRTF banks being implemented in the application:

1. Default Unity HRTF bank (no additional information is available for the public)
2. Standard HRTF of humans – ARI – NH72 (*hrtf_nh72.sofa* [59]) as *HRTF_set_1*
3. HRTF of an artificial head – ARI (ARTIFICIAL) – measured on dummy head Neumann KU 100, NH172 (*hrtf_b_nh172.sofa* [60]) as *HRTF_set_2*

Both downloaded HRTF banks were measured with the resolution of 2.5° within the ± 45° range, and with a 5° resolution beyond this angle. The measurements were evenly distributed across the sphere, yielding 1550 HRTFs in one set [61]. Unfortunately, anthropometric data were marked down only for the first 60 targets (all available here [62]), so the details for HRTF sets downloaded by me are not available. Even though it is highly inconvenient, it should not pose a problem for the final results evaluation, as the main goal was to test for perceivable differences and for possibility of future HRTF testing.

On an important note, both HRTFs were measured as far field HRTFs (see *2.4.1. HRTF, ITD, ILD*). Because there is a big difference between near and far field HRTFs, it is important to place the detectable objects and distractors in greater distance than one meter away from the player - marginal distance between near and far field (counting in the possibility that the testee might move inside the player area set up by the HTC Vive bases and unintentionally get closer).

Audio tracks were exchanged as well. Following the previous conclusion, music is not suitable for this type of task and in order to maintain as much ecological validity as possible, I wanted to avoid pulses, constant frequency sounds and alike. Also, too complex sounds might be harder to visually represent a confuse the subjects. To fit the environment, created in the first stage, natural sounds would be appropriate and choice comes down to which. Forests and nature seem appropriate for bird chirping and singing. As distractors, other nature found sounds were chosen. All of them described in the following paragraph.

Four different bird chirps and songs were used, in order to avoid subject irritation and include more variety to the experiment. All of the four audio tracks are of high quality and sufficient for audio quality testing. As distractors, a forest stream, rain, an angry squirrel and otter squeaking were found to fit the environment very well and all of them are easily distinguished from the target audio source – bird chirping. All the audio tracks were download and are credited. Additionally, some were slightly adjusted to better fit the test. Specifications for every sound are listed below:

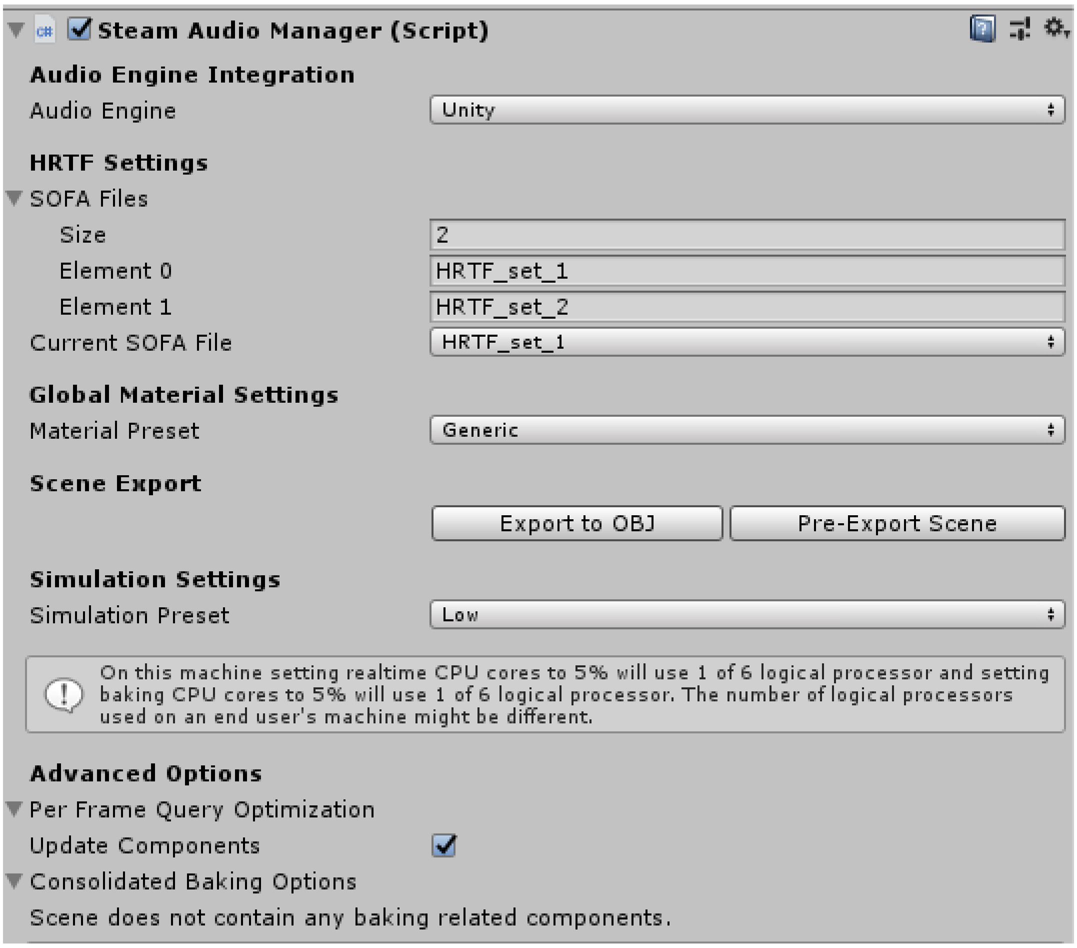
- Bird chirping [63] [64] [65] [66] – 3x 44 100 Hz, 1x 48 000 Hz, 16-bit, 2x 24-bit, 32-bit, wav format
- Forest stream [67] – 44 100Hz, 24-bit, wav format
- Rain [68] – 48 000 Hz, 24-bit, wav format
- Otters squeaking [69] – 44 100 Hz, 24-bit, wav format
- Squirrel chatter [70] – 44 100 Hz, mp3 format

With new audio, graphics needed to be updated as well. As the concept of the final test design is the effects of distractors, a visual representation of distractors and audio sources had to be created. Initially, all the objects were represented by white spheres of various sizes and changed color on controller point. However, to increase immersion and the overall feeling of authenticity, spheres were replaced by real graphic content. In correspondence with chosen sounds, I deemed birds to be the most appropriate representation and as mentioned earlier, Unity Assets store offers many free packages. One of these is *Living Birds* [71] package. *Living birds* features many kinds of birds, which were ideal for the sound source and distractor representation and also allowed to input a variety to the test (not just one object to represent everything). Some other distractors were used as well and for these *Simplistic Low Poly Nature* [72] package was downloaded and applied. Squirrel and turtle objects were picked to represent different kinds of distractors.

3.3.2. NEW TEST DESIGN

The environment altogether remains the same. Natural scenery proved to be a suitable environment and provided sufficient immersion. It will also correspond with newly chosen audio tracks and their representing objects.

Functions of this test concept, however, will be much more extensive in comparison to the previous version. Additionally, some new objects had to be put in place inside the scene itself. Every audio source had to be given a new component from the Steam Audio Plugin. *Steam Audio Source* adds new possibilities to *Audio Source* components (mentioned in *2.3.1. Steam Audio Plugin*). As a separate game object, *Steam Audio Manager* has to be created as it handles the HRTF banks and corresponding settings which need to be partly setup prior the test run. Even though the *Steam Audio Manager* allows for exchanging HRTFs on runtime, they need to be specifically setup beforehand. Figure 18 shows such configuration. The number of SOFA files and their exact names have to be set. HRTF banks, with these names, are then copied inside the *Streaming Assets* folder and can afterwards be accessed via script/command on application runtime.



☑ Steam Audio Manager (Script)

Audio Engine Integration

Audio Engine: Unity

HRTF Settings

SOFA Files

Size: 2

Element 0: HRTF_set_1

Element 1: HRTF_set_2

Current SOFA File: HRTF_set_1

Global Material Settings

Material Preset: Generic

Scene Export

Simulation Settings

Simulation Preset: Low

On this machine setting realtime CPU cores to 5% will use 1 of 6 logical processor and setting baking CPU cores to 5% will use 1 of 6 logical processor. The number of logical processors used on an end user's machine might be different.

Advanced Options

Per Frame Query Optimization

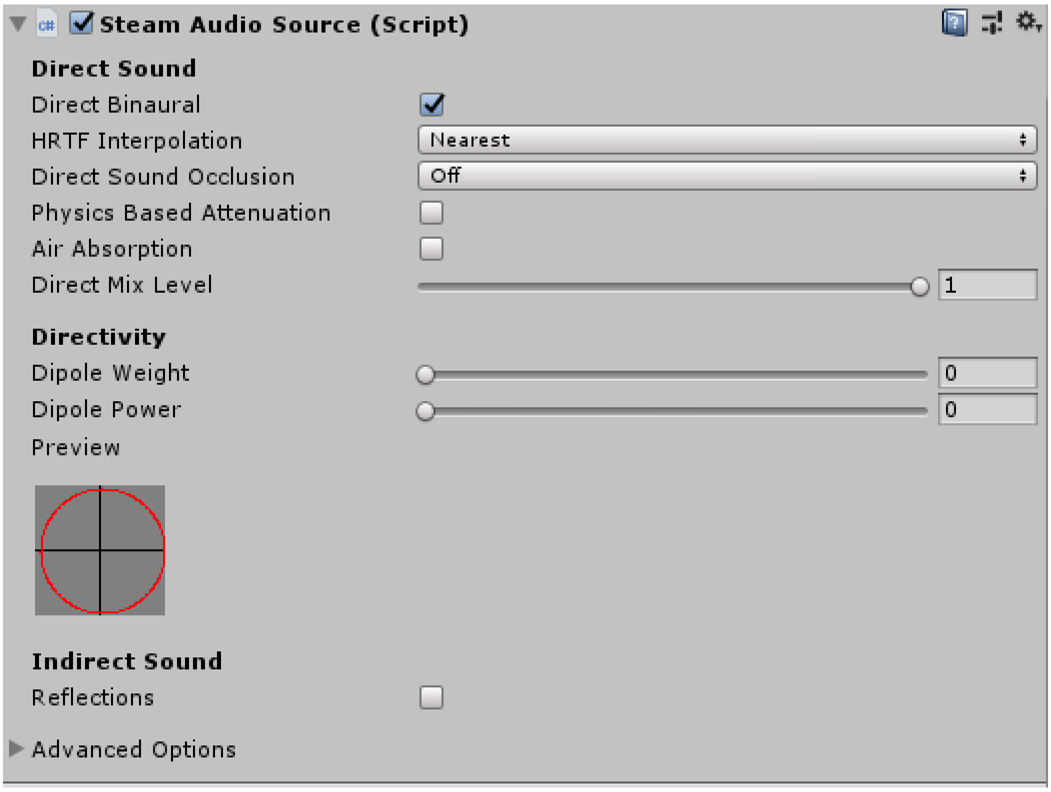
Update Components ☑

Consolidated Baking Options

Scene does not contain any baking related components.

Figure 18: Steam Audio Manager component

Steam Audio Source (Figure 19) controls soundwave's behavior and the way the sound is rendered. Without this component, it would not be possible to imitate interaction between the soundwave and the environment. Also, the interpolation selection is set inside this component and determines, how the renderer handles missing HRTF direction.



☑ **Steam Audio Source (Script)**

Direct Sound

Direct Binaural ☑

HRTF Interpolation: Nearest

Direct Sound Occlusion: Off

Physics Based Attenuation □

Air Absorption □

Direct Mix Level: 1

Directivity

Dipole Weight: 0

Dipole Power: 0

Preview

Indirect Sound

Reflections □

▶ Advanced Options

Figure 19: Steam Audio Source component

In this final test, more information needs to be marked down before starting the test session. Because of that, four different initial input fields were added. All are launched in succession. Once the value of the previous one is confirmed (using *Enter* key), the next one pops up. The original bug, where once clicked outside the field, the field could no longer be filled, was fixed, and now clicking away from the field allows for complete overwrite of the text inside the field. Placeholders were put in place to navigate the expected input. The four windows are:

1. Name input
2. Sensory impairment information
3. Selection of the type of interpolation
4. Selection of desired HRTF bank

Name input and sensory information fields take in any value and save it to the *Results.txt* text file inside the *Streaming Assets* folder. However, that is not the case of the next two fields. The next is the interpolation selection field, where the input of "1" selects bilinear interpolation (= interpolation of the 4 nearest functions) and every other input selects nearest interpolation. The last input is the HRTF selection. This design supports 2 additional HRTF banks in SOFA, poetically named *HRTF_set_1* and *HRTF_set_2*, which need to be located inside the *Streaming Assets* folder with this exact name. Their respective numbers select them and any other input selects Unity's default HRTF. Depending on the choices in the HRTF and interpolation fields, the application correspondently applies these settings inside the test. HRTF set is switched inside the *Steam Audio Manager* and can be then observed as "*Current SOFA file*" (Figure 18).

The implementation of the input fields can be seen in the Figure 20 below.

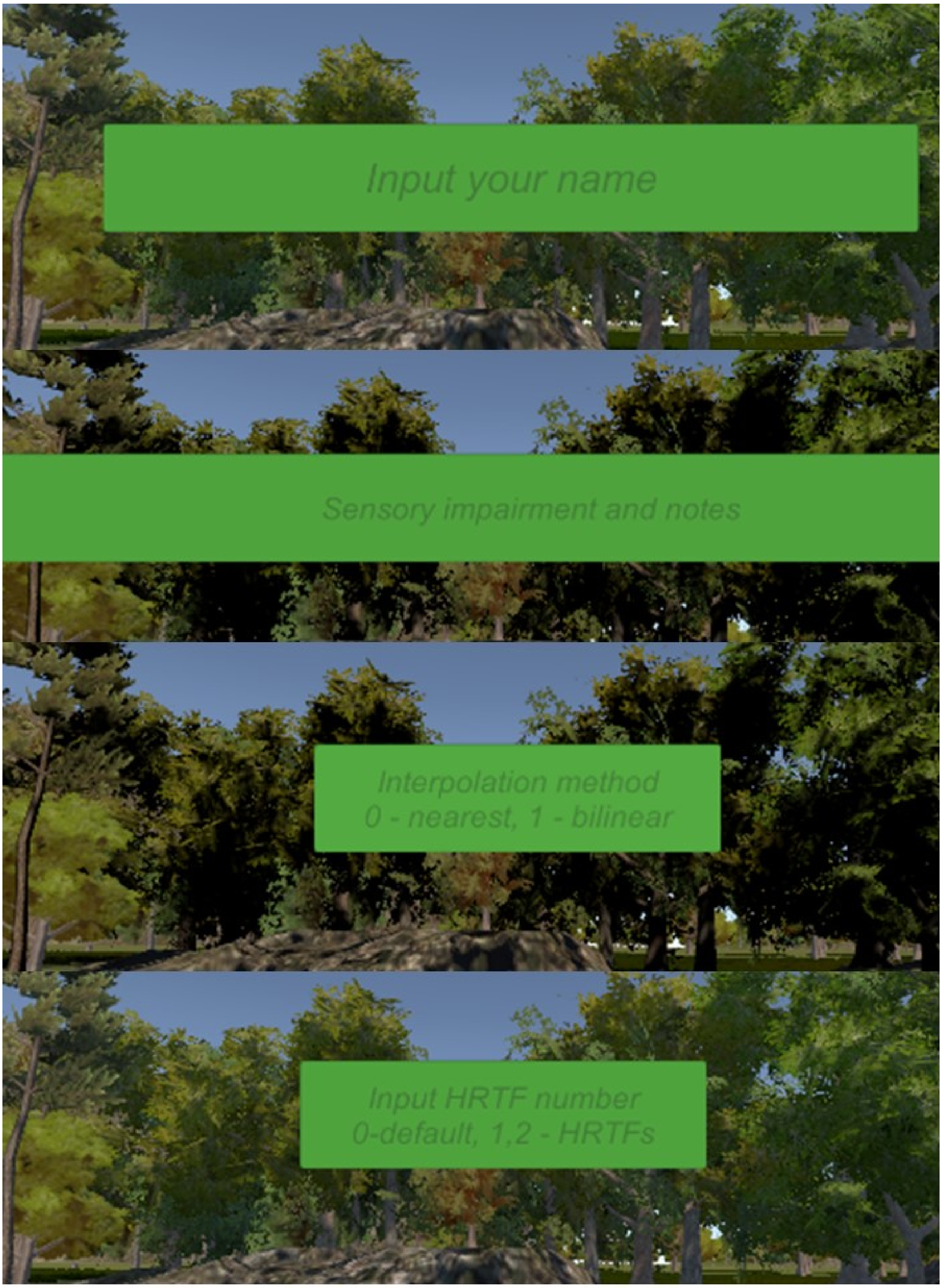
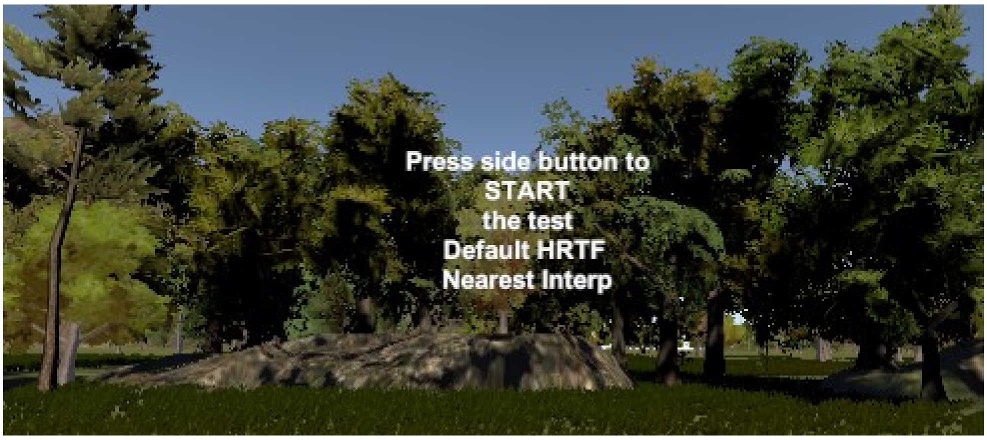


Figure 20: Implementation of input fields

Choices of HRTF set and interpolation method are written in the results text file and are summed up in the prompted text (Figure 21) that appears after confirming all input fields and asks the testee to launch the test session.



Press side button to
START
the test
Default HRTF
Nearest Interp

Figure 21: Prompted start text with HRTF and interpolation overview

Afterwards, the test is launched and the test duration counter is started. Termination of this counter is realized with tagging the last audio source, in test case number 31.

Selecting the presumed audio source realization remains the same as before, using a raycast accompanied by laser pointer. In order to make the selected object clearly visible, there is an invisible sphere around every selectable object, which changes its color on raycast hit. In this case, it changes to translucent green. The realization example can be seen in Figure 22. These spheres also register the raycast hit on selection and mark the target as the picked sound source. Even though the colliders around the confined player area are still in place, in this test setting, they do not register hits by the raycast line and therefore the only way to advance to the next test case is selecting a valid object.

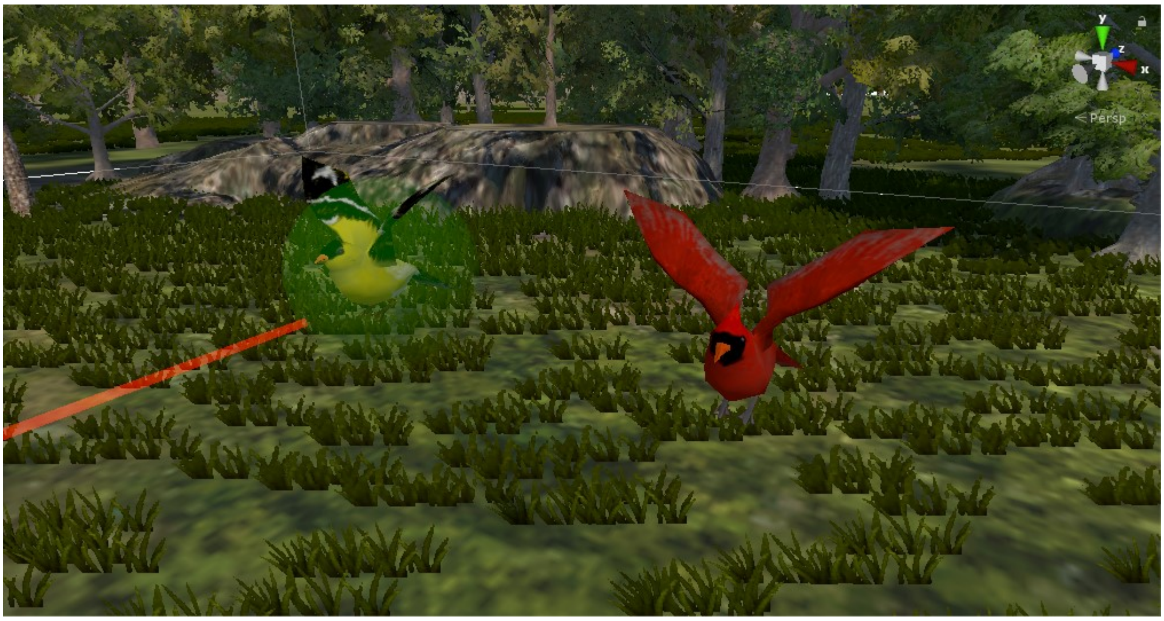


Figure 22: Raycast tagging execution example

As mentioned in the previous subchapter, execution of the visual representation of the sound source object is done using prefabbed bird models. In every test case, the spheres were replaced by various types of birds. That provides more distinguishability between single objects and adds more believability. Their positions are precalculated and fixed (more on this topic will be written in subchapter *3.3.4. Test cases*). And with changing distance, the size had to be often adjusted as well, which sometimes leads to worsened visibility. The object formations vary, but Figure 23 displays and example of object setup.

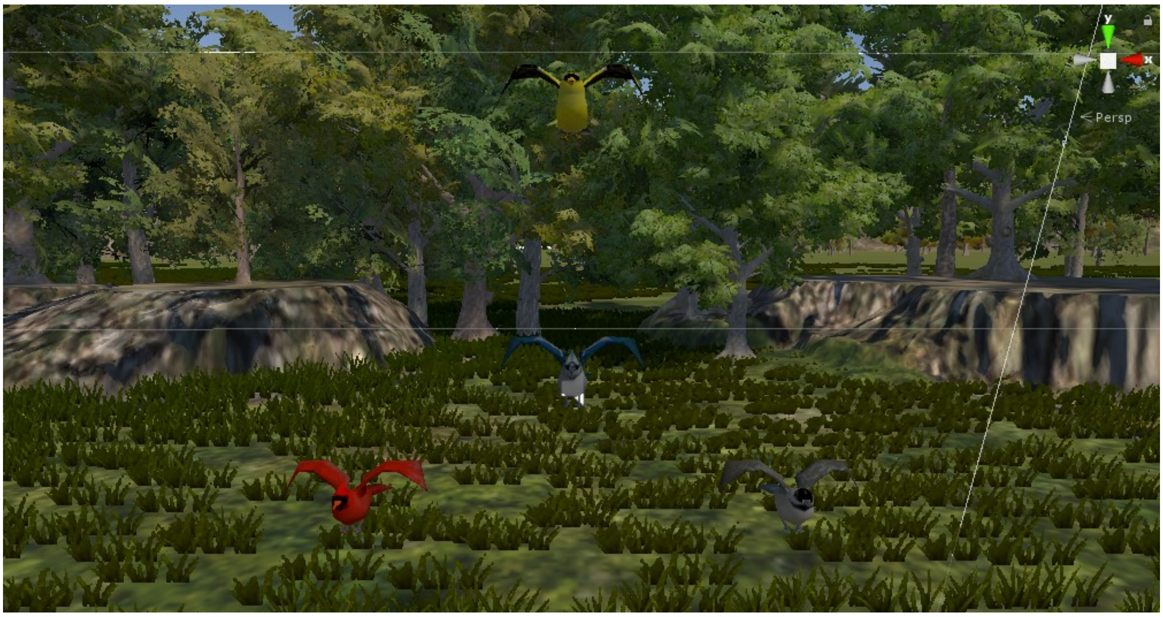


Figure 23: Visual representation of possible sources - formation example

Different object scales were sometimes part of the test specification, and therefore led to different-sized source representations – Figure 24.

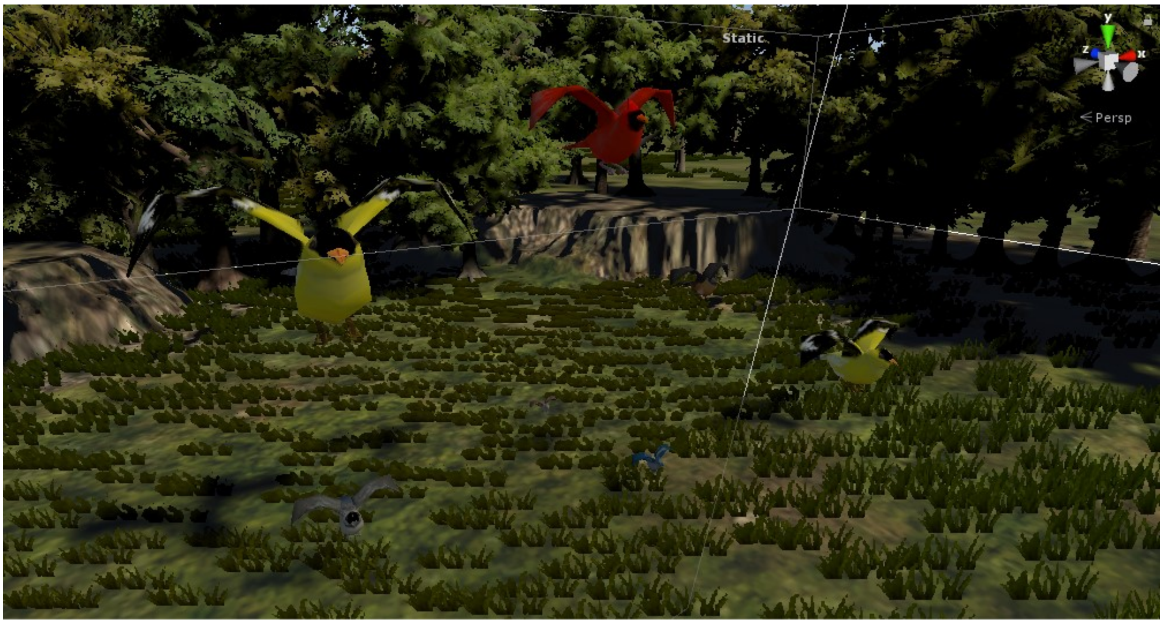


Figure 24: Visual representation of possible sources - size variation

Apart from regular sources, distractors were implemented as well and in order to only test the effect of a distractor as a whole, a different visual representation (as well as sound) was chosen for the distracting source. The next picture (Figure 25) displays setup using a squirrel distractor.

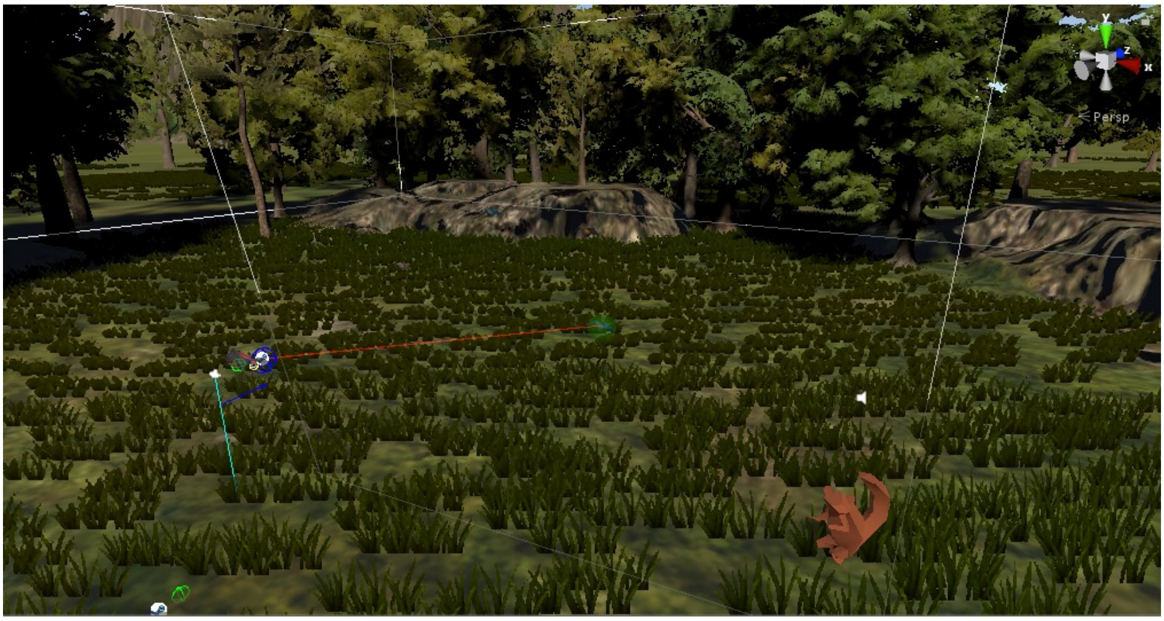


Figure 25: Distractor implementation

The following picture (Figure 26), just to see the concept, depicts the initial state of the test session. All of the test cases are overlapping right before the application is launched (and disabled on startup).

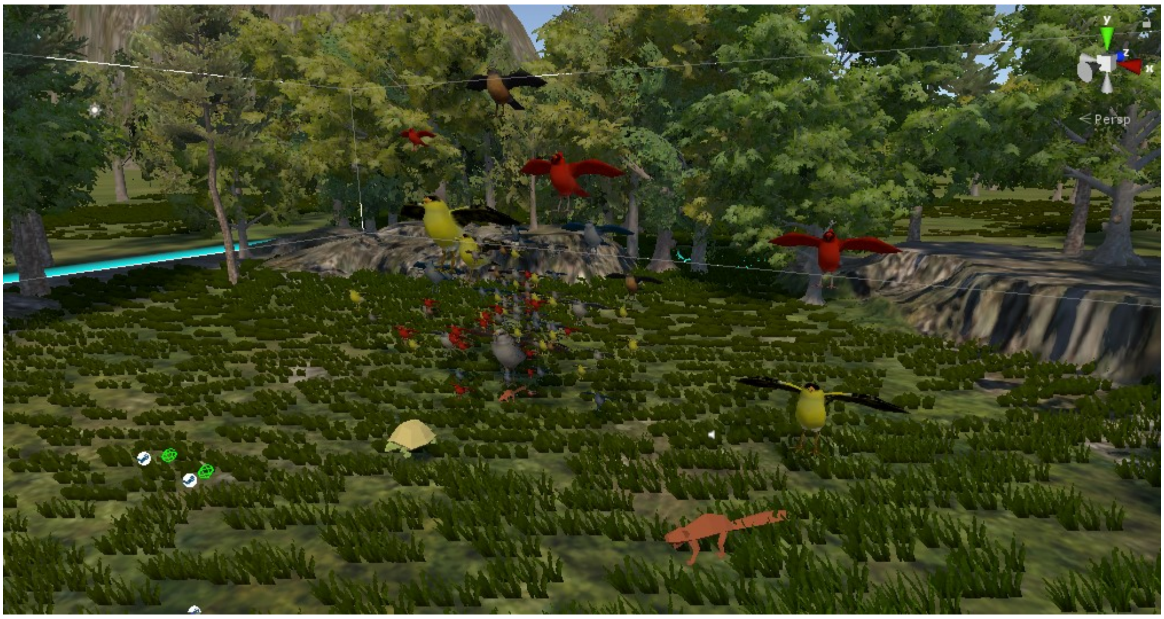
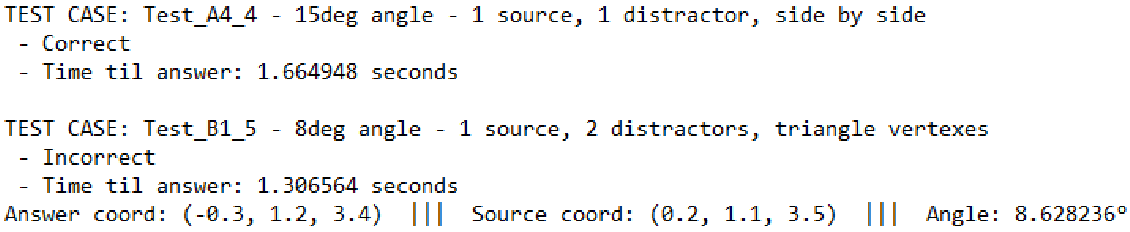


Figure 26: All test cases overlapping overview

Apart from writing down the initial settings and information (name, sensory imp., interpolation method and HRTF], every test case is also marked down. Every test case is fully described by the test name and its description (which is taken from the description element on every case). Furthermore, whether the answer was correct or not and the reaction/decision time. In case the answer was correct, only this fact is noted. However, if the answer is incorrect, the system will also find the location of the presumed source, the real audio source and calculates the angle going towards the object from the player position (assumes the position [0,0,0]). It also parses the value into degrees as Unity's standard is radians. The text output can be seen below in Figure 27.



```
TEST CASE: Test_A4_4 - 15deg angle - 1 source, 1 distractor, side by side
 - Correct
 - Time til answer: 1.664948 seconds

TEST CASE: Test_B1_5 - 8deg angle - 1 source, 2 distractors, triangle vertexes
 - Incorrect
 - Time til answer: 1.306564 seconds
Answer coord: (-0.3, 1.2, 3.4)  |||  Source coord: (0.2, 1.1, 3.5)  |||  Angle: 8.628236°
```

Figure 27: Output formatting example

The location of the *Results.txt* file is in the *Streaming Assets* folder inside the *Build* folder. Usually, the *Streaming Assets* folder is also present in the *Assets* folder in the project, but this has been setup simply to avoid having to change the path in case the application is built to an .exe file in the future.

After every source selection, the next test case is started and the previous one disabled. When all of the test cases are finished, a new text appears and asks the testee to take of the headset (Figure 28). After that the test can be terminated via Unity interface by the overseeing person.

Figure 28: Final screen – test finished

3.3.3. Test flow

Test flow, in general, remains the same as in the previous stage. The overseeing person instructs the subject. The testee has the HMD, headphones and controller comfortably put on. The overseeing person types in subject's name, sensory impairments, chooses the interpolation method and the HRTF set. Afterwards, the user presses the side button on the controller and the test begins.

Reaction/decision time is marked down for every test, but the subject cannot see it. The only thing visible to the user is the test itself and the environment, nothing apart from that and there is no feedback regarding the answer. The testee points the controller at the presumed sound source, the object is highlighted (see Figure 22) and using the mechanical click of the trigger, the selection is marked down. After every selection, values are sent to the text file and the next test case is launched, up until all of the 31 cases are depleted.

At the end, a prompted text announces the end of the test and asks the subject to take of the headset. All of the answers can then be found in the *Results.txt* file inside the *Streaming Assets* folder inside the *Build* folder. The format of the answers can be seen in Figure 27 above. Finally, the supervisor shuts down the Unity project and everything is automatically ready to be launched again.

3.3.4. TEST CASES

This section describes the composition of test cases. They are thematically comprised, going from simpler cases up to more complex concepts. All the cases consist of distractor exploration, most of them purely visual, but in some of them, audio distractors are added as well. In order to have full control, there is no ambient sound, despite the environment being a nature scenery.

Every test consists of the target audio source, distractors (disguised as possible audio sources), distractor audio sources, which are always different to the target sound to avoid unnecessary confusions and a description that can be seen from inside Unity inspector and is marked along with the test case answer.

It is important to note that all the values in programmed test cases are inspired by studies in *2.8. Background research on VAS audio quality and HRTF based measurements*. The player is in the center of the playing space, therefore the [0,0,0] coordinates. The distance between the objects representing sound sources is precisely calculated to contain angles 3°, 5° and 8° as per MAA in [18]. And after the initial testing, a 15° case was added as well. Firstly, to compare it to to the other cases and secondly, to cover the error from [29]. These together should be on the edge of human perception, and should therefore be ideal for researching the effect of distractors.

As the distance between two separate sources could always be somehow calculated as a triangle, with the wanted angle clenched between two sides at the player location, all the calculations were done using variations of goniometric functions and triangle properties. Considering it is not possible to always maintain the same angle between all of the objects, some compromises had to be made.

Test cases numbered from 1 to 16 consist of four different cases, featuring various numbers of objects with the distance matching the 3°, 5°, 8° and 15° angles. The layout can be seen in the Figure 29 below, where *X* represents the calculated distance.

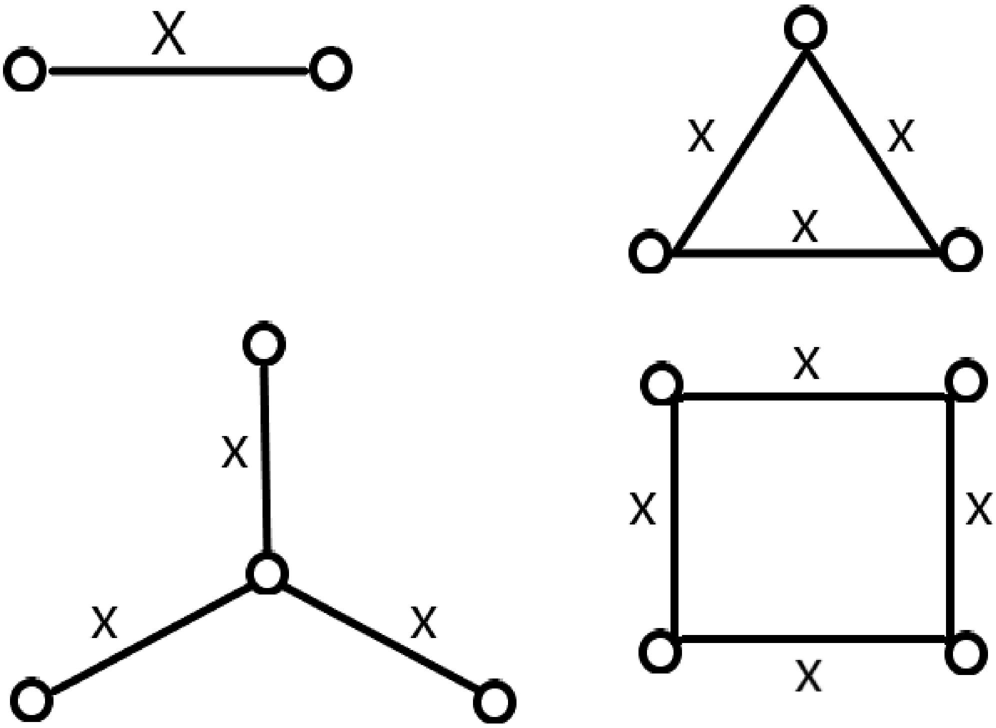


Figure 29: Formation of possible sources - test cases 1-16

Cases from 17 to 24 show application of various movement, which differs in axes and restrictions:

- 17 and 18 form a square, with different object spacing that corresponds to angles of 8 and 15 degrees, with movement restricted to X axis.
- 19 consists of four objects, but the movement is restricted to X axis and randomized. The same goes for test number 20, but the number of objects is increased to seven.
- Case number 21 returns to square formation, with 15° angle corresponding spacing and movement restricted to XY axes. 22 already has an increased count of objects to 7 and randomized XY movement.
- Numbers 23 and 24 copy the same concept as latter cases, but with omnidirectional movement

The next three, numbers 25, 26 and 27 explore the influence of size of the object on testee's decision. 25 consists of four objects, 26 of seven objects, all with static positions. Test case 27 expands to 10 objects and adds omnidirectional movement to some of them.

Four test cases that follow are all different:

- Number 28 consists of 10 visual objects, with added circling distractors – they have different diameters and speed, but the true audio source remains static.
- Case 29 features 4 objects plus 2 distractor audio sources. One in the form of an animal and the other one a flowing spring. Both of them static.
- Test case number 30 uses 10 objects, including two audio distractors as well, but adds movement. Both of them are circling around the subject, while the true audio source remains static.

- The final test case contains 8 objects, plus the sound of rain and its visual representation as well.

All of these always remain the same in every test run. It is for the purpose that certain positions might be easier to detect, as they often differ in both horizontal position and vertical position. And because there is no immediate feedback and the tested subject never finds out the correct answers, it does not influence the test results even while tested multiple times with different HRTFs.

3.4. FINAL TESTING

Final testing was conducted on 10 different subjects. Three of them with background in acoustics and music, three of them ballroom dancers, which gives them deeper musical feeling and four of them with no musical background at all. Only the three subjects with the background in acoustics had previous experience with audio testing. Apart from that, just one out of the 10 subjects had previously worked with virtual reality. As the test does not contain a training session for inexperienced users, nor does it give feedback on the answers, I decided to use the test itself as a training environment. Testees were allowed to move around and try the virtual environment prior starting the test. It was either in the Steam VR Home or in the test itself, because they were not informed of their choices regarding the audio sources, so it could not have affected the final testing.

Additionally, sensory impairment information was collected for every subject in order to conclude, whether a different set of results might have been affected by it. The average age of the tested people was 23 years and comprised of both males and females alike.

Moreover, two of those subjects were given two consequential tests, with a 5-minute break in between, each one with a different HRTF set to compare, whether a difference will be heard or whether it affects the results in any way (this also partly proves that the application is capable of changing HRTF banks on runtime).

Specifications of the testing configuration correspond to an above average gaming system. As mentioned in [53], ecological validity is an important aspect of VR experiments and using special equipment, e.g. professional sound cards or top of the line headphones, would not simulate regular conditions under which VR headsets are being used. The specifications are as follow:

- Win 10, Intel i5 8000 series, Nvidia GTX 1070 Ti graphics card, 32 GB RAM
- Circumaural gaming headset GX Zabius HS-G850, 20 Hz – 20 kHz, 32 Ohm impedance, sensitivity dB – SPL 117
- HTC Vive HMD (*3.1.1. Head Mounted Display Selection*)

All tests were conducted with the same equipment, same setup and the same volume settings. The experiment room was calm and quiet, with no ambient noise and no low bass noise (caused by for example elevators, adjacent streets with busy traffic and so on). Subjects were seated in front of the computer and instructed to not move around the playing area, but were free to move and rotate their head without restrictions.

Finally, four different parameters were marked down:

1. Correct/incorrect answer
2. Time from the start of the test case to selection of presumed audio source
3. Time it has taken to complete the entire test
4. Angle between two sides of a triangle, going from the player location to selected and presumed audio source

4. Results and discussion

Final testing results are displayed in Table 3 below. Detailed results for every subject are available in Appendix B on the enclosed CD.

Table 3: Table of results from the final testing phase

TEST	DESCRIPTION	CORRECT ANS	AVG RT [s]	AVG ANG ERROR [°]
1	8deg angle - 1 source, 1 distractor, side by side	9	10.19823345	7.328747
2	5deg angle - 1 source, 1 distractor, side by side	4	8.794711917	5.118430375
3	3deg angle - 1 source, 1 distractor, side by side	7	14.73526867	2.987076
4	15deg angle - 1 source, 1 distractor, side by side	11	8.203247583	14.98264
5	8deg angle - 1 source, 2 distractors, triangle vertexes	8	17.40810908	7.02993025
6	5deg angle - 1 source, 2 distractors, triangle vertexes	4	14.706706	4.2517075
7	3deg angle - 1 source, 2 distractors, triangle vertexes	2	11.42570158	2.8970861
8	15deg angle - 1 source, 2 distractors, triangle vertexes	11	12.181794	14.26764
9	8deg angle - 1 source, 3 distractors, triangle vertexes + center	8	18.67614883	7.430903
10	5deg angle - 1 source, 3 distractors, triangle vertexes + center	5	12.53842267	7.795714857
11	3deg angle - 1 source, 3 distractors, triangle vertexes + center	9	7.620563	4.077879667
12	15deg angle - 1 source, 3 distractors, triangle vertexes + center	4	20.04026858	18.03825625
13	8deg angle - 1 source, 3 distractors, square	5	12.04772792	7.424809714
14	5deg angle - 1 source, 3 distractors, square	4	12.16198483	4.846964375
15	3deg angle - 1 source, 3 distractors, square	4	13.56411475	3.134091375
16	15deg angle - 1 source, 3 distractors, square	7	20.04052225	13.956932
17	8deg - 1 source, 3 distractors, specified X movement, square	7	23.10830333	7.0001422
18	15deg - 1 source, 3 distractors, specified X movement, square	6	15.39805025	13.57097417
19	15deg ,1 source, 3 distractors, random X movement	9	19.60918942	7.307581667
20	1 source, 6 distractors, random X movement	8	19.13406375	14.2358425
21	15deg, 1 source, 3 distractors, specified XY movement, square	10	19.7494885	10.567385
22	1 source, 6 distractors, random XY movement	8	22.65055033	12.3444655
23	1 source, 3 distractors, specified omnidirectional movement, square	9	13.70010233	9.638379333
24	1 source, 6 distractors, random omnidirectional movement	10	18.95217475	15.9231
25	1 source + 3 distractors - The influence of source size 1	6	17.97880275	15.55665667
26	1 source + 6 distractors - The influence of source size 2	9	9.394540667	18.95507
27	1 source + 8 distractors + moving - The influence of source size 3	10	11.629386	27.316065
28	1 source, 9 distractors, circling distractors	5	14.97046967	14.71877714
29	3 sources, 3 distractors, 2 static distractor audios	11	20.33291	16.09152
30	3 sources, 7 distractors, moving distractor audio	6	23.51084275	11.4869575
31	2 sources, 6 distractors, static distractor audio behind the player, with rain	9	12.9389985	10.989047

The average angular error was calculated only from the incorrect answers, in order to portray the average indistinguishability. Red background symbolizes that the average angle error is calculated from only one or two values and therefore does not carry much weight.

Reason, why the erroneous angles are not entirely corresponding with the calculated angles of 3°, 5°, 8° and 15° is that even though the player teleportation was not allowed, player still could have moved in the real-world space, which might have caused inaccuracy.

The results, in conjunction with the positions of various audio sources, also clearly show that often the subject recognizes the correct side of the object formation, where the sound is coming from, but struggles with deciding the correct object in terms of elevation. That can be seen in Table 3 above. Tests numbered 4 and 8, featuring 15-degree angle, have the lowest error rate out of the first 8 tests. Tests numbered from 9 to 16, are more dependent on vertical recognition, which is obviously weaker link in the human perception. Test cases 17-31 are an overview of all possible tests and are hard to evaluate as there are always only a few cases of each type.

However, drawing from the study of Tew and Kelly [18], regarding the angles of 3° and 5°. These are evidently not applicable in real applications as human perception is not that sensitive (their test study was quite specific and the type of detection is not really applicable in regular VR applications).

Test cases which added movement indicate that when the motion was restricted to one axis, it was harder to determine, which audio source is the correct one, while more degrees of freedom made the task easier. Here, however, also come into question the speed of the objects. During the initial testing, it seemed that higher speed led to easier distinguishability, but made the marking impossible, because the birds became untargetable at such speed. Therefore, the subject stopped paying attention to the task and focused on targeting instead. Slower speed made the targeting easier, but sometimes led to more overlapping of objects, which in turn concluded in more difficult recognition. For that reason, both of these traits need to be kept in balance.

On the other hand, there are many test cases which could lead to separate testing. For example, the influence of size regarding the decision making. Expectation is that people would tend to choose bigger sources over smaller ones. As the smallest audio source is in test case number 25, which has the highest error rate, it suggests some truth to that assumption.

The average reaction time corresponds with Fang Chen's experiment [29] and his discoveries. Localization time mentioned in his study was 14.7 ± 9.8 seconds for untrained targets. Considering not all test cases in my experiment were ideal (clearly detectable), the time still corresponds with his findings. Moreover, there is no trend in shorter time to number of correct answers per test ratio.

Only two of the three HRTF sets were tested, therefore the third HRTF bank (see *3.3.1. Used assets,* ARI artificial) will be available inside the application for future work. The Default Unity HRTF and the human ARI HRTF. Two subjects performed consequential tests on Default HRTF, following with the human ARI HRTF. Both subjects noted that a difference is perceivable and that the human ARI HRTF provides clearer sound.

Interestingly, even though the final number of correct answers (almost) does not differ, the answers that were correct in one test were not correct in the next one and vice versa. This points to audible differences in those two HRTFs and therefore also proves that this application can provide feedback on various HRTF banks.

4.1. Unity as a tool and HMD evaluation

Unity, version 2018.2.6f1, proved to be dependable platform and fully suitable for this assignment. It offers wide variety of assets and tools, which in connection to SteamVR, allow for flexible application structure that can be shaped accordingly to accommodate different needs. It also enables quick and simple changes to the test design which allow for swift adjustments and open this test to future manipulation and optimization.

Using Steam Audio Plugin allowed for HRTF exchanging and different interpolation methods to be implemented inside the test and proved to be easy to use and more than suitable for this task.

Chosen HMD, HTC Vive, appears sufficient for this assignment. The graphics did not disturb the overall feeling of immersion, the 3.5mm jack allowed connection of different headphones and controls are simple and intuitive. For future testing, headset with eye tracking might provide different possibilities and deeper insight, but it did not affect this particular experiment.

It was also decided, to leave the application in a Unity project state and not built it at the moment. Unity project allows fast adjustments and changes depending on desired test cases and outcomes and leaves the project open to quick manipulation.

5. Conclusion

Background research regarding VAS, audio quality testing in virtual reality, VR development platforms, HMDs and their supported environments was conducted in the extent covering everything needed for creating a fully functional experimental application. A conclusion was drawn from various studies and researches and the topic of distractors was chosen in order to take a more specific approach towards VAS audio quality testing.

The application itself was created in Unity Engine environment and developed for HTC Vive headset, using SteamVR platform. Unity proved to be a fitting tool and HTC Vive was deemed a suitable choice.

A first pilot test was created and tested in the terms of immersion, environment, controls, content and conclusions for the next test structure. Following up on findings from the first test design, another test architecture was devised and programmed to feature test cases to test distractors in AV integrated scene. Extensional functions were implemented in order to add the possibility to choose interpolation method for rendering missing direction functions and to change HRTFs on runtime. Additionally, the first design was published as a contribution during the 23rd International Student Conference on Electrical Engineering, Poster 2019.

Even though the results are not fully conclusive, and it is not possible to create a statistically supported conclusion, the test confirmed some hypotheses. It also disproves the MAA angles discovered during background research for regular VR use. The results also show that there is a clear connection between perceived audio location and distractors and a lot of potential for future studies.

Moreover, the application can switch between various HRTFs on runtime and despite the fact that it is not possible to compare used HRTFs (Default Unity HRTF does not have available parameter specifications and neither does the second HRTF), differences were observed. The oral evaluation as well as the results show that there is distinguishable variation.

The application as a whole fulfills the desired parameters and is proven to be reliably working and no bugs or crashes were encountered during the final testing. The final stage of the application remains as a Unity project to allow easy access and adjustments. The application measures decision time of every test, global test duration, correct/incorrect answer and calculates erroneous angle between the answer and the correct source, contained from the player position.

The test structure allows easy alterations and is opened to customization in the future. More research into HRTF comparison and into various aspects of distractors using this application could be conducted. The source code, along with the Unity project are attached as Appendix A, and may serve as basis for future research.

BIBLIOGRAPHY

[1] Staff Writer, "BusinessTech," BusinessTech, 14 10 2017. [Online]. Available: https://businesstech.co.za/news/technology/204358/the-future-of-tv-is-mobile-on-demand-and-heading-to-virtual-reality/. [Accessed 9 2 2019].

[2] D. Sedláček, *3D Modeling and Virtual Reality Presentations,* Prague: CTU University subject, DCGI, 2018.

[3] Blizzard team, "World of Warcraft," Blizzard, [Online]. Available: https://worldofwarcraft.com/en-us/. [Accessed 3 1 2019].

[4] S. Shahrbanian, X. Ma, N. Aghaei, N. Korner-Bitensky, K. Moshiri and M. J. Simmonds, "Use of virtual reality (immersive vs. non immersive) for pain management in children and adults: A systematic review of evidence from randomized controlled trials," *European Journal of Experimental Biology,* vol. 2, no. 5, pp. 1408-1422, 2012.

[5] V. S. Editor, "Virtual Reality Immersion," Virtual Reality Society, 2017. [Online]. Available: https://www.vrs.org.uk/virtual-reality/immersion.html. [Accessed 3 13 2019].

[6] FeelReal Team, "Feelreal Sensory Mask," FeelReal, Inc., 2018. [Online]. Available: https://feelreal.com/. [Accessed 12 4 2019].

[7] T. Bordwell and M. Reize, "Learning Space dedicated to the Art and Analyses of Film Sound Design," FilmSound.org, [Online]. Available: http://filmsound.org/terminology/diegetic.htm. [Accessed 11 1 2019].

[8] "Unity," [Online]. Available: https://unity3d.com/. [Accessed 2 7 2018].

[9] "Unreal Engine," [Online]. Available: https://www.unrealengine.com/en-US/what-is-unreal-engine-4. [Accessed 2 7 2018].

[10] I. Dudkin, "UNREAL VS UNITY FOR VR DEVELOPMENT," Skywell Software, 21 1 2019. [Online]. Available: https://skywell.software/blog/unreal-vs-unity-for-vr-development/. [Accessed 14 3 2019].

[11] "Steam Audio," [Online]. Available: https://valvesoftware.github.io/steam-audio/. [Accessed 7 10 2018].

[12] S. Editor, "SOFA (Spatially Oriented Format for Acoustics)," Sofa Conventions, 2018. [Online]. Available: https://www.sofaconventions.org/mediawiki/index.php/SOFA_(Spatially_Oriented_Format_for_Acoustics). [Accessed 12 2 2019].

[13] S. Carlile, Virtual Auditory Space: Generation and Applications, Berlin: Springer, ISBN 978-3-662-22594-3, 1996.

[14] D. Rao, B.-S. Xie and G.-Z. Yu, "Characteristics of Near-Field Head-Related Transfer Function for KEMAR," in *40th International Conference: Spatial Audio: Sense the Sound of Space*, Tokyo, Japan, 2010.

[15] L. Betbeder, "Near-field 3D Audio Explained," 19 9 2017. [Online]. Available: https://developer.oculus.com/blog/near-field-3d-audio-explained/. [Accessed 5 4 2019].

[16] F. L. Wightman and D. J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America,* vol. 105, no. 5, pp. 2841-1853, 1999.

[17] A. W. Mills, "On the Minimum Audible Angle," *The Journal of the Acoustical Society of America,* vol. 30, pp. 237-246, 1958.

[18] M. C. Kelly and A. I. Tew, "A novel method for the efficient comparison," in *AES, 114th Convention, Convention Paper 5786*, Amsterdam, The Netherlands, 2003.

[19] T. Kuppanda et. al., "Virtual Reality Platform for Sonification Evaluation," in *The 21st International Conference on Auditory Display (ICAD 2015)*, Graz - Austria, 2015.

[20] J.-M. Pernaux, M. Emerit and N. Rozenn, "Perceptual Evaluation of Binaural Sound Synthesis: The Problem of Reporting Localization Judgements," in *AES 114th Convention*, Amsterdam, The Netherlands, 2003.

[21] D. Poirier-Quinot and B. F. Katz, "Impact of HRTF individualization on player performance in a VR shooter game I," in *Conference on Spatial Reproduction*, Tokyo, Japan, 2018.

[22] D. Poirier-Quinot and B. F. Katz, "Impact of HRTF individualization on player performance in a VR shooter game II," in *Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, USA, 2018.

[23] Author Unknown, "Comparing methods of testing - subjective methods," Unviersity of Salford, Manchester, [Online]. Available: https://www.salford.ac.uk/research/sirc/research-groups/acoustics/psychoacoustics/sound-quality-making-products-sound-better/accordion/sound-quality-testing/sound-quality-testing-subjective-methods. [Accessed 19 4 2019].

[24] J.-M. Pernaux, M. Emerit, D. Jerome and N. Rozenn, "Perceptual Evaluation of Static Bunaural Sound Synthesis," in *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002.

[25] V. Larcher, in *Techniques de spatialisation des sons pour la réalité virtuelle*, Paris, Université de Paris VI, 2001, p. 195.

[26] M. C. Kelly and A. I. Tew, "The significance of spectral overlap in multiple-source localization," in *AES 114th convention*, Amsterdam, The Netherlands, Convention Paper 5725, 2003.

[27] O. Rummukainen, S. J. Schlecht, A. Plinge and E. A. P. Habets, "Evaluating binaural reproduction systems from behavioral patterns in a virtual reality - A case study with impaired binaural cues and tracking latency," in *AES 143rd convention*, New York - USA, 2017.

[28] J. S. Calle and A. Roginska, "Head Rotation Data Exttraction From Virtual Reality Gameplay Using Non-Individualized HRTF," in *AES 143rd convention*, New York - USA, 2017.

[29] F. Chen, "The Reaction Time for Subjects to Localize 3D Sound via Headphones," in *AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio.*, Espoo - Finland, 2002.

[30] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio and Z. Zhang, "Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity," *Frontiers in Neuroscience,* vol. 12, no. DOI=10.3389/fnins.2018.00021, p. 21, 2018.

[31] H. Møller and e. al., "Head-related transfer functions of human subjects.," *Journal of the Audio Engineering Society,* vol. 5, no. 43, pp. 300-321, 1995.

[32] P. Majdak, R. Baumgartner and B. Laback, "Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization," *Frontiers in psychology,* vol. 5, p. 319, 2014.

[33] A. Unkown, "Sofa Conventions," The Acoustics Research Institute of the Austrian Academy of Sciences, 5 9 2018. [Online]. Available: https://www.sofaconventions.org/mediawiki/index.php/Files. [Accessed 12 3 2019].

[34] I. B. Witten and E. I. Knudsen, "Why seeing is believing: merging auditory and visual worlds.," *Neuron ,* vol. 48, no. 3, pp. 489-496, 2005.

[35] A. A. Ghazanfar and C. E. Schroeder, "Is neocortex essentially multisensory?," *Trends in cognitive sciences,* vol. 10, no. 6, pp. 278-285, 2006.

[36] B. E. Stein and T. R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron.," *Nature Reviews Neuroscience,* vol. 9, no. 4, p. 255, 2008.

[37] K. Connolly, "Multisensory perception as an associative learning process," *Frontiers in*

psychology, vol. 5, p. 1095, 2014.

[38] E. Paraskevopoulos and e. al., "Evidence for training-induced plasticity in multisensory brain structures: an MEG study.," *PloS one,* vol. 7, no. 5, p. DOI 0036534, 2012.

[39] B. Bonath and e. al., "Neural basis of the ventriloquist illusion.," *Current Biology ,* vol. 17, no. 19, pp. 1697-1703, 2007.

[40] G. H. Recanzone, "Rapidly induced auditory plasticity: the ventriloquism aftereffect," *Proceedings of the National Academy of Sciences,* vol. 95, no. 3, pp. 869-875, 1998.

[41] I. Frissen, J. Vroomen and B. de Gelder, "The aftereffects of ventriloquism: the time course of the visual recalibration of auditory localization.," *Seeing and perceiving,* vol. 25, no. 1, pp. 1-14, 2012.

[42] I. Frissen and e. al., "The aftereffects of ventriloquism: generalization across sound-frequencies," *Acta psychologica,* vol. 118, no. 1-2, pp. 93-100, 2005.

[43] N. Kitagawa and S. Ichihara, "Hearing visual motion in depth," *Nature,* vol. 416, no. 6877, p. 172, 2002.

[44] C. C. Berger, Ehrsson and H. Henrik, "Auditory motion elicits a visual motion aftereffect," *Frontiers in Neuroscience,* vol. 10, p. 559, 2016.

[45] J. H. McDermott, "The cocktail party problem," *Current Biology,* vol. 19, no. 22, pp. R1024-R1027, 2009.

[46] O. Rummukainen, J. Wang, Z. Li, T. Robotham, Z. Yan, Z. Li, X. Xie, F. Nagel and E. A. P. Habets, "Influence of visual content on the perceived audio quality in virutal reality," in *AES 145th Convention, Convention Paper 10128*, New York, 2018.

[47] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 364, no. 1535, pp. 3549-3557, 2009.

[48] U. Reiter and M. Weitzel, "Influence of interaction on perceived quality in audiovisual applications: evaluation of cross-modal influence," in *13th International Conference on Auditory Display (ICAD)*, Montreal, Canada, 2007.

[49] D. Thery and e. al., "Impact of the Visual Rendering System on Subjective Auralization Assessment in VR," in *Virtual Reality and Augmented Reality*, Springer, 2017, pp. 105-118.

[50] D. Alais and D. Burr, "The ventriloquist effectresults from near-optimal bimodal integration," *Current Biology,* vol. 14, no. 3, pp. 257-262, 2004.

[51] O. Rummukainen and C. Mendonça, "Task-relevant spatialized auditory cues enhance atten-tion orientation and peripheral target detectionin natural scenes," *ournal of Eye Movement Re-search,* vol. 9, no. 4, pp. 1-10, 2016.

[52] D. C. Niehorster, L. Li and M. Lappe, "The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research," *Iperception, DOI: 10.1177/2041669517708205,* p. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5439658/, 18. 5. 2012.

[53] B. Olk, A. Dinu, D. J. Zielinski and R. Kopper, "Measuring visual search and distraction in immersive virtual reality," *Royal Society Open Science (Volume 5, Issue 5),* vol. 5, no. 5, 2018, May.

[54] "VIVE™ | Buy VIVE Hardware," Vive, [Online]. Available: https://www.vive.com/eu/product/#vive-spec. [Accessed 30 01 2019].

[55] "SteamVR™ Tracking," [Online]. Available: https://partner.steamgames.com/vrlicensing. [Accessed 30 01 2019].

[56] Valve Corporation, *SteamVR Plugin,* Valve Corporation, Jan 2019.

[57] Unity Technologies, *Standard Assets,* Unity Technologies, Mar 2018.

[58] Author Unknown, "Files," SofaConventions, 2019. [Online]. Available: https://www.sofaconventions.org/mediawiki/index.php/Files. [Accessed 15 4 2019].

[59] Austrian Academy of Sciences, *hrtf_nh72.sofa,* HRTF: http://sofacoustics.org/data/database/ari/.

[60] Austrian Academy of Sciences, *hrtf b_nh172.sofa,* HRTF: http://sofacoustics.org/data/database/ari%20(artificial)/.

[61] Author Unknown, "Detailed Description," Austrian Academy of Sciences, [Online]. Available: https://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=721&catid=165&lang=en&Itemid=670. [Accessed 10 5 2019].

[62] Author Unknown, "ARI HRTF Database," Austrian Academy of Sciences, [Online]. Available: https://www.kfs.oeaw.ac.at/index.php?view=article&id=608&lang=en#AnthropometricData. [Accessed 5 5 2019].

[63] InspectorJ, *Bird Whistling, A.wav,* Audio Track: https://freesound.org/people/InspectorJ/sounds/339326/.

[64] InspectorJ, *Bird Whistling, Robin, Single, 13.wav,* Audio Track: https://freesound.org/people/InspectorJ/sounds/456440/.

[65] InspectorJ, *Bird Whistling, Single, Robin, A.wav,* Audio Track: https://freesound.org/people/InspectorJ/sounds/416529/.

[66] Lizardhood, *Birds Chirping.wav,* Audio Track: https://freesound.org/people/Lizardhood/sounds/427040/.

[67] odilonmarcenaro, *forest stream close recording,* Audio Track: https://freesound.org/people/odilonmarcenaro/sounds/235943/.

[68] tim.kahn, *light forest rain.wav,* Audio Track: https://freesound.org/people/tim.kahn/sounds/169031/.

[69] O. Squeaking.wav, *Motion_S,,* Audio Track: https://freesound.org/people/Motion_S/sounds/221761/.

[70] bmccoy2, *Squirrel Chatter 4 3 2016 Lincoln Nebraska .mp3,* Audio Track: https://freesound.org/people/bmccoy2/sounds/342105/.

[71] Dinopunch, *Living Birds,* Assets Package, 2014.

[72] Acorn Bringer, *Simplistic Low Poly Nature,* Assets Package, 2018.

LIST OF FIGURES

LIST OF TABLES

APPENDIX Ax (On CD)

Appendix A contains every part of the Unity project

- **A1** DP_VAS_Test_Using_HMD - folder containing the entire Unity project
- **A2** Source_Code.zip - .zip folder with sources codes
- **A3** Virtual_Acoustic_Space_Test_Using_HMD_Instructions.pdf – PDF file with instructions for future test adjustments

APPENDIX Bx (On CD)

Appendix B contains everything from the final testing stage

- **B1** RESULTS.zip – All 12 .txt test output files
- **B2** Test_results_processing.xlsx – Excel with processed test data
- **B3** TrianglesCalculations.m – Objects distance calculations