UNIVERSITY OF TARTU VILJANDI CULTURAL ACADEMY
Faculty of Arts and Humanities
Sound and Visual Technology

Viljar Rosin

PUULUUP IMMERSIVE

Masters project

Supervisors:

Christoph Felix Schulz

Allowed for defence ……..

Janar Paeglis

Allowed for defence ……..

Jürgen Volmer

Allowed for defence ……..

Viljandi 2021

TABLE OF CONTENTS

1. INTRODUCTION

The aim of this project is to use different spatial audio recording and mixing techniques to give the listener the feeling of being surrounded by nature or the environment where the visual content is taking place and music concurrently – a musical short movie that can be viewed anywhere with the feeling of being in the open air, on the street or even under water. The usage of surround sound in music has been growing throughout the recent years. There are even different albums available on Spotify that have been mixed for binaural, Ambisonic or surround listening. The aim for creating such content is to give the listener a feeling of being immersed and inside the music and/or environment. “Everyday life is full of three-dimensional sound experience.“(Rumsey, 2001, p. 1)

Leaving the experience to be not only imaginable, there will be also visual content with different experimantal variations of video capturing technicues. Now, it is very common for home producers to film 3D videos with the equipment that is available inexpensivly from the internet. When talking about 3D, one can think instantly of 3D animation motion pictures, but this project’s aim is to use stereoscopic video shooting with two small sports cameras (GoPro) and fuse footage into one using the method of Anaglyph 3D. “Anaglyph stereo provides a low-budget solution to viewing stereoscopic images. However, it may suffer from ghosting and bad color reproduction.”(Sanftmann & Weiskopf, 2011, p. 1251) Anaglyph 3D viewing requires red-cyan glasses, which might be familiar from different movies where the audience is wearing cardboard/paper glasses, and a screen (a TV or a computer). For these reasons, anaglyph seemed the best option for the project due to its inexpensivness while also creating visiual spatiousness.

For creating the content, the band Puuluup will be participating in the audio recording and filming process. While the music will be recorded in a controlled environment as a live performance, the video will be shot outside and that makes the recording of audio more intricate because of the noises that will be produced by the crew (footsteps, talking etc.) or unexpected noises that must be minimalized as much as possible. For this purpose, the technique of Foley will be used which means actually re-recording the scenery sounds with an Ambisonic 1st order microphone. Finally, it will be decoded into a binaural and 5.1 surround sound format.

2. SURROUND SOUND & 3D SOUND

We are surrounded by different sounds everywhere and through the diffusion and reflections of sound waves, we gain perspective. In everyday life we don't think about hearing as a mechanism of time and space perception. In the real world, audio engineers are always trying to solve the problems of how to reproduce the real-life feeling and perception with technical equipment. "Auditory perception is a complex phenomenon determined by the physiology of the auditory system and affected by cognitive processes. The auditory system transforms the fundamental independent aspects of sound stimuli, such as their spectral content, temporal properties and location in space into distinct patterns of neural activity. These patterns will give rise to the qualitative experience of pitch, loudness, timbre and location."(Roginska & Geluso, 2018, p. 5)

Surround sound can be divided into two categories: 2D and 3D. 2D surround sound means that all the sound reproducing objects have been placed mainly on the same level. This means the listener will be surrounded by sound from all around. Sound waves reflect on physical objects, which may trick the listener into hearing some sounds coming from above or below. 3D on the other hand has been developed the way that there are ceiling speakers and sometimes also floor or "underground" speakers that provide the immersive experience of being surrounded by sound from every possible angle. The "underground" speakers are uncommon, although are integrated into some of the 3D surround setups.

3. INTRODUCTION TO SPATIAL AUDIO

The reproduction of sound has evolved over time from one sound source to the point, where a sound system can be istalled and feature loudspeakers at almost every figurable place – the floor (or under the seats), the walls and the ceiling. In this next section there will be a brief overview of the different types of configurations. In the book "Instant Surround" the author, J. P. Fisher, explains different ways of sound reproduction in a simple way by starting with one source and building up to surround. The following chapters will provide an overview of the most common reproduction types of sound.

3.1 Mono

In mono (monaural/monophonic), meaning "one", a single speaker is needed for reproducing sound. The audio that is being played back does not have to be recorded with only one microphone or reproduced by one speaker, but the sound/music that is sent to the speakers is excactly the same. It is a common practice for audio mixing engineers to check their mixes in mono to see if there are phase cancellations or other problems when listening to the mix.

3.2 Stereo

The most common audio reproduction and format in everyday life is stereo (stereophonic) which means the sound will be presented by two speakers within a stereo track. The audio sources can be placed within the stereo field. Listening to audio in the "sweet spot" means both speaker are located

at approximately 30 plus or minus 10 degrees from the center and equidistant from the listener. although, it is apparent that the sound only comes from the front. “The stereo experience has a sense of depth and space to it, with many sounds seemingly emanating from the centre even though no speaker is actually there. This phenomen is called the phantom center.“ (Fisher, 2005, p.8)

Most music that we listen to is mixed in stereo. Conventionally, we listen to music with a 2-speaker setup (i.e. speakers, headphones, car speakers, television, etc.) common headphones consist of two drivers – one for the left and one for the right ear resulting ultimately in stereo. Even though in the 1970s the quadrophonic home systems became available for the common listener, the stereo system still had the upper hand and stayed as a market leader throughout that time. “Stereo continues to be the main format for music reproduction, having survived while other more complex systems have not. It is a relatively simple system and gives a good impression of spaciousness, suitable for enjoyable music listening. In short; overall, it is the best solution.” (Hallum, 2017, p. 10)

3.3 3.0

Adding one more speaker to the stereo is called a 3.0 system, that means a center speaker is added between the stereo mode that is called the center speaker. The film industry typically utilizes this system. Sound designers conventionally place the dialog is in the center speaker and the ambience in the two side speakers to create space and directionality. “Adding a center channel (3.0 or C) to the stereo pair was a Hollywood invention to solve the problem of being outside the stereo sweetspot.“ (Fisher, 2005, p. 9)

3.4 Quadrophonic - 4.0

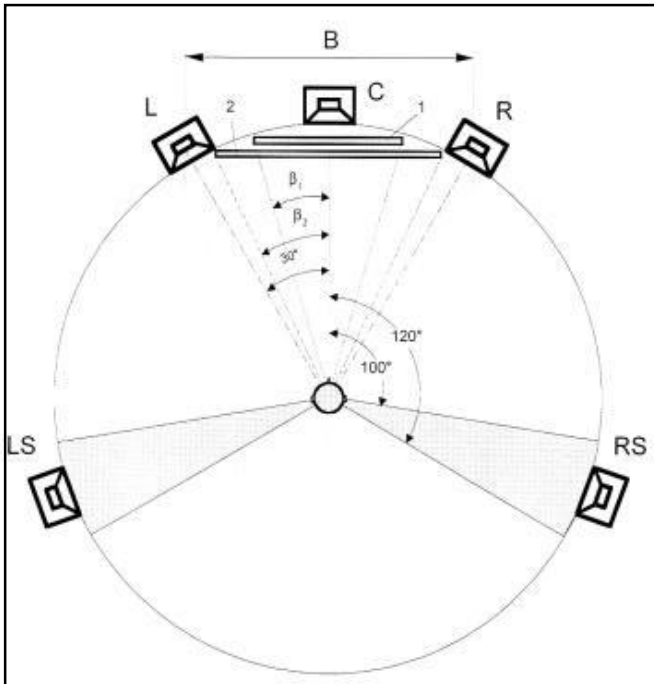
An uncommon set up still used today is quadrophonic (4.0) – Left/Right+rear Left/Right. The usage of quadrophonic recording and reproduction became extinct for many different reasons. According to the article “Competing Networks and Proprietary Standards: The Case of Quadraphonic Sound“ by

Steven R. Postrel in 1990, fatal for the quadrophonic system became the value, the lack of demand, the size of the user base, underpromotion and the release date of the system. “The costs of producing four-channel records instead of two-channel were substantial. Producers, mixers, and artists had to make decisions about where to locate performers and microphones in the studio, and how to mix down the many recorded tracks into four channels to create the desired sound image.“(ibid, p. 182)

Even though it was a big financial decision, still many records were released in quadrophonic, for example on the 1st of March in 1973 Pink Floyd’s album, *The Dark Side of the Moon*, was released on a vinyl as quadrophonic recording. 1974 Aerosmith released their album, *Get Your Wings*, and on the same year The Allman Brothers released, *Eat a Peach*.

3.5 5.1 and beyond

The most popular home-theater system is a 5.1 system – L/C/R + rear L/R and a subwoofer is added (as the x.**1** applies). This setup became really popular for home-theater systems in the 1995 with the arrival of DVDs (digital versatile disc) which featured the possibility of compressing and encoding audio tracks in Dolby AC-3 – the DTS (Digital Theatre Systems) and Dolby Digital Sound. Recommendation for setting up the 5.1 system by Audio Engineering Society, INC (see figure 1.1).



Screen 1: Listening distance = 3*H* (2ß1 = 33°)

Screen 2: Listening distance = 2*H* (2ß2 = 48°)

H: Screen height

B: Loudspeaker basis width

* Depending on shape, type, and size of screen.

Acoustical Center	**Angle**	**Height**	**Tilt**
C	0°	1.2m*	0°*
L, R	±30°	1.2m	0°
LS, RS	±100–120°	³1.2m	£ 15°

Figure 1.1 Reference loudspeaker setup with loudspeakers L/C/R and LS/RS, in combination with picture reproduction installation (in accordance with ITU-R BS. 775-1) (Rumsey et al, 2001, p. 4)

The 7.1 system consists of the setup of 5.1 system and there are additional side speakers – side left and side right. The setup can be adapted accordingly to the needs, depending on the reproducible variables, it can be from 5.1 up to infinity.1. This system has been successfully adopted by the music and film industry and the most used codecs for these applications are made by Dolby.

When talking about immersive sound or surround sound systems usually Dolby is mentioned more than once, even movies and video games promote Dolby – it has become a standard in this field. There are different types of setups for this system and the most known and used one is Dolby Atmos.

3.6 Dolby Atmos

This system has its own software for creating spaciousness and it requires specifically designed speakers for Dolby Atmos. These speakers are equipped with multiple drivers, some of them are facing up, which actually use the ceiling for reflecting sounds from above. For the best listening experience it is necessary to install all the speakers at the correct distance, angle and placement from the "sweet spot". The instructions can be found on the Dolby website.

> "Dolby Atmos is based on the concept of sound objects. In the cinema, Dolby Atmos relies on a combination of 9.1 ''bed'' channels and up to 118 simultaneous sound objects to deliver an enveloping sound stage. Every sound in a scene – a child yelling, a helicopter taking off, a car horn blaring – can be a separate sound object. Each of those sounds comes from a specific location in the scene, and in some cases, they move. The car careens from left to right, while the yelling child runs up a set of stairs." (Dolby Atmos for the Home Theater, 2016, p. 3)

Dolby describes the different variations on their website as the following:

"Dolby Atmos speaker layouts parallel the 5.1 and 7.1 setups for surround sound.

- A 5.1.2 or 7.1.2 system uses two ceiling speakers, or two Dolby Atmos-enabled speakers or modules.
- A 5.1.4 or 7.1.4 system uses four ceiling speakers, or four Dolby Atmos-enabled speakers or modules.
- A 9.1.2 system adds a pair of front wide speakers to a 7.1.2 layout."

(Dolby Atmos Speaker Setup...2021)

There are also other companies who have “improved” the surround sound experience, but they are not so widely used or known: Auro - 3D, 360 RA (Sony 360 Reality Audio) etc.

3.7 The Binaural system

“Human hearing is capable of selecting single sounds from a mixture of sounds while suppressing the unwanted components (the cocktail party effect). This is done in the listener’s brain by exploiting the ear signals as two spatially separated sound receivers in a process frequently referred to as *binaural signal processing*.” (Ballou, 2008, p. 553)

Headphones are the most commonly used “sound system“ in everyday life. Listeners are listening to music at public places, transportation, and even at home. This system can be hands-free, bluetooth, and even with built-in microphones. While the common format is stereo, binaural has been gaining popularity. For binaural listening a pair of headphones or just two speakers is needed, but listening in an open space, the reflections will create a misunderstanding of the spatialization and speakers don’t provide accurate representations of binaural space.

> “The perception of binaural sound relies on Interaural Time Difference (ITD) and Interaural Intensity Difference[1] (IID) cues. Together, the ITD and IID are the foundation of Lord Rayleigh’s Duplex theory (1907). In addition to the interaural cues, spectral information and variations as a function of location provide invaluable information to the listener about the position of a sound source. The composite of the ITD, IID and the spectral coloration characteristics are captured in Head-Related Transfer Functions (HRTF).” (Wenzel, 1992, cited via Roginska & Geluso, 2018, p. 88)

The purpose of the binaural sound reproduction is to create a real-life like listening experience, where the aim is to create 3D surround for headphones. It is very common to render Ambisonic recordings to binaural with the help of different softwares and plugins. Although there are many benefits to

binaural recording and mixing, the technology is still evolving. “One major problem of binaural hearing is that for high spatial localization accuracy, the HRTFs must be measured for the targeted human head and for different azimuth and elevation angles.” (Tsakostas et al., 2007, p. 292) Within certain binaural decoding plugins there are default presets, that have been measured for certain types of human head and ear size and spacing.

3.7.1 Binaural microphones

Since the binaural microphone consists of two, as identical as possible, capsules, it is sometimes classified as a stereo microphone. “The final stereo microphone to be considered is not really a stereo microphone at all, but rather a binaural microphone called a dummy head. Stereo is distinguished from binaural by stereo being aimed at loudspeaker reproduction, and binaural at headphone reproduction. Binaural recording involves a model of the human head, with outer ears (pinna), and microphones placed either at the outer tip of the simulated ear canal, or terminating the inside end of an artifi cial ear canal. With signals from the microphones supplied over headphones, a more or less complete model of the external parts of the human hearing system is produced.” (Holman, 2008, p. 85–86)

Binaural microphones are designed to be as close as possible to the human ear. One example of human head like microphone design for recording in binaural is the Neumann KU100 (see figure 2.1) - it is called “The Dummy Head“. The Dummy Head is a replication of the human head with ears and it uses diffuse-field equalization for stereophonic recordings, which means, that listening the recording via headphones should give you the feeling of being positioned exactly at the same place, where the Dummy Head was positioned during the recording. The microphone is also used for documentation of noise at different halls/work places/etc.

Figure 2.1 Dummy Head KU-100 (Source: Neumann 2021)

3.8 Ambisonics

Ambisonic technology has been around for some time although has garnered popularity amongst Virtual reality (VR) in the recent years. Because of this popularity, Ambisonics has continued to progress while the industry is also making equipment available to common users at an affordable pricepoint.

> "Ambisonics is a method of codifying a sound field taking into account its directional properties. In traditional multichannel audio (e.g., stereo, 5.1 and 7.1 surround) each channel has the signal corresponding to a given loudspeaker. Instead, in Ambisonics each channel has information about certain physical properties of the acoustic field, such as the pressure or the acoustic velocity.
>
> Ambisonics is a perturbative theory:
>
> 0. At *zeroth order*, Ambisonics has information about the pressure field at the origin (recording of an omnidirectional microphone at the origin). The channel for the pressure Field is conventionally called *W*.
>
> 1. At first order, Ambisonics adds information about the acoustic velocity at the origin (recording of three figure-of-eight microphones at the origin, along each one of the axis). These channels are called *X, Y, Z*. Following the Euler equation, the velocity vector is proportional (up to some equalization) to the gradient of the pressure field along each one of the axis.
>
> 2. At second and higher orders, Ambisonics adds information about higher order derivatives of the pressure field.
>
> In this course we will limit ourselves mostly to first order Ambisonics, which we shall call simply Ambisonics from here on, although we will also do a brief introduction to higher order Ambisonics (HOA)."(Arteaga, 2015, p. 4)

The technology of sound recording and reproduction has evolved to the possibility of 3D audio, which creates the illusion of directionality, depth, and spatial recognition. The stereo system mainly uses the X and Y axis, to provide information of directinality, either the sound is coming from the left, right or from the center. In this system the Z axis is also used for depth and it shows us, if the sound is far away or really near. When it comes to surround sound, the Y-axis is added to create the feeling of height to make the experience more real – for instance a plane is flying overhead of the listener or a bird is singing on top of a tree.

> "Using an approach known as Ambisonics B-format[34] the sound information is encoded into four separate channels labeled W, X, Y and Z. The W channel would be equivalent to the mono output from an omnidirectional microphone while the X, Y and Z channels are the directional components of the sound in front-back (X), left-right (Y), and up-down (Z) directions. This allows a single B-format file to be stored for each location to account for all head motions at this specific location and to produce a realistic and fast auralization as the user can move from one receiver location to the other and experience a near-seamless simulation even while turning his/her head in the virtual model." (Ballou, 2008, p. 235)

The A-format in Ambisonics refers to the raw recording. If the audio is already decoded into a surround format then it is called B-format. Within Ambisonic audio there is a hierarchy known as "orders". The order indicates the number of channels that are being recorded. Afterwards when it comes to decoding and reproducing, the order shows how many channels are available for use.

> "While even first-order B-format provides higher-resolution spatial immersion than traditional surround technologies, higher-order B-format audio can provide even higher spatial resolutions, with more channels providing more different polar patterns. Thus, second-order Ambisonics uses 9 channels, third-order Ambisonics jumps up to 16 channels, all the way up to sixth-order Ambisonics with 49 channels."(*Ambisonics Explained*...2017)

For recording Ambisonic audio, it is imperative that all the channels that are being used have the same gain structure while also being in the correct sequential order for the next process: decoding. In the case of gain differences, the decoding creates a false image, because the formulas will miscalculate the image and depth.

> "There are two methods to process Ambisonic parameters: one is matrix inversion, another one is vector resultant, based on psychological acoustics: virtual sound synthesis. The precision of Ambisonic technology depends on the installation requirements. The more loudspeakers are used in playback, the more real audiences will percept. Ambisonic technology cannot totally eliminate the "sweet spot". If the playback system is in a higher standard, the scope of "sweet spot" will be larger."(Lu, 2015, p. 51)

3.8.1 Ambisonic microphones

Ambisonic microphones are referred to as first order Ambisonics as they record the raw uncoded audio. The build of Ambisonic microphones is different depending on what is needed. Usually the basic microphone is presented with 4 matched capsules. "An Ambisonics recording microphone is built of four microphone capsules encased closely together. These capsules are cardioid polar patterns, and the signals they record are usually referred to as "Ambisonics A-format." The A-format is then transformed to B-format by a simple matrix to the WXYZ channels."(*Ambisonics Explained*...2017). For example Sennheiser AMBEO VR-microphone, it has 4 outputs, but for example Zoom H3-VR has already a builtin decoder. There are also more complex and expensive Ambisonic microphones that consist of up to 32 omni capsules; some may have even more. These microphones require decoding through software and the more capsules on the microphone, will require more free space on your computer for recording. Since these microphones are meant to record spatial audio, manufacturers mostly provide free decoders with the microphone.

3.9 Microphone techniques for spatial audio

In the field of sound recording and reproducing, a huge leap was made in 1876-1877 when Edison created a cylinder-based phonograph, which was a monophonic sound reproducer. Since then, we have reached the era of astonishing technical advancements, where immersive audio is apart of our lives with the help of home-theater systems and specific surround systems at different art exhibitions, cinema and theater halls.

While working with spatial audio, it is necessary to know the desired deliverable for the listener or the type of speaker setup the listener is meant to listen with. The conversion from surround to mono/stereo is a common process, but vice versa is possible only when someone has the access to the primary multi-track recording. There is also the possibility to fake surround with plug-ins that will create a false feeling of spatial panning.

Mono and stereo recording techniques are the fundamentals for surround sound, when it comes to the practice of each microphone being routed via a matrix to a certain single speaker, this is and has been one of the practices of surround arrangements. For the 5.1 system for example, the Decca tree technique with back-facing two microphones can be used succesfully for recording ambience and for the 0.1 the low end will be taken from an omni microphone, which provides the low end. There is also the method of using many spot microphones and afterwards in the post production the mixing engineer will arrange/mix the spot microphones accordingly for the reproducable system to make the listener feel like one is in the middle of an orchestra or on a street corner. ”Microphone setups for multichannel include use of standard microphones in particular setup combinations, and microphone systems designed as a whole for multichannel sound.”(Holman, 2008, p. 71)

Depending on the final result, the speaker setup and location, spatial recording techniques may differ. One must always keep in mind that recording audio has many proven techniques, but there is also a lot of space for improvisation depending on source. Recording music or ambience have different common techniques and no recording situation or environment is exactly identical. Commonly used

and simple surround microphone techniques use the most known microphone placements – XY, M/S, Blumelin and ORTF. “XY postitions a matched pair of cardioids close together but pointed in opposite directions, 90 degrees away from center. ORTF, positions two matched cardioids, pointed in a V pattern 17cm and 110 degrees apart.” (Fisher, 2005, p. 26)

When choosing the right microphones for recording spatial audio, one must choose between different types of microphones to achieve the desired result. “Omnidirectional mics provide better low frequencies, and a very immersive, enveloping image. Cardioid and hyper-cardioid mics provide better channel separation, more open, and potentially more accurate imaging. (...) In music production there are no specifc “.1” mics. Low frequency content is directed to the LFE channel from the main mics, or any spot mics on sound sources which have important low frequency content.”(Corbett, 2021, p. 163)

For stereo recording there are also special microphones that already consist of a matched pair of capsules and many of the modern handheld recorders already have a matched pair of capsules attached to them – for example portable recorders by Zoom.

There is also the Decca Tree option which is referred to as a stereo microphone technique, but it consists of 3 microphones – LCR. For 5.1 recording there will be added two back-facing microphones to the Decca Tree, so it will be L/C/R/Ls/Rs.

After recording all the material, the audio has to be decoded for accurate special image and rendered for reproduction.

3.10 Audio reprodution and rendering

Mostly, the rendering of spatial audio is data-based or model-based and the reproduction and mixing is channel-based, object-based or scene-based. These methods can be combined with eachother, depending on what is needed.

> "There are a number of different ways to represent a soundfield. Eample formats include channel-based audio formats, object-based audio formats and scene-based audio formats. Channel-based audio formats refer to the 5.1 surround sound format, 7.1 surround sound formats, 22.2 surround sound formats, or any other channel based format that localizes audio channels to particular locations around the listener in order to recreate a soundfield. (...) Object-based audio formats may refer to formats in which audio objects, often encoded using pulse-code modulation (PCM) and referred to as PCM audio objects, are specified in order to represent the soundfield. Such audio objects may include metadata identifying a location of the audio object relative to a listener or other point of reference in the soundfield, such that the audio object may be rendered to one or more speaker channels for playback in an effort to recreate the soundfield. The techniques described in this disclosure may apply to any of the foregoing formats, including scene-based audio formats, channel-based audio formats, object-based audio formats, or any combination thereof." (Kim & Peters, 2020, p. 2)

3.11 Mixing immersive audio

Surround sound is all about panning the sounds into different sound sources and there are two main methods for that. "Panning is used in two senses: fixed assignment of microphone channels to one or more loudspeaker channels, called static panning, and motion of sound sources during mixing, called dynamic panning. Of the two, static panning is practiced on nearly every channel every day, while dynamic panning is practiced for most program material much less frequently, if at all."(Holman, 2008, p. 115)

Mixing audio that has been recorded with an Ambisonic microphone is all about decoding the A-Format to the right B-Format Higher Order Ambisonics and with the process it is possible to rotate the audio as needed for the project that is in progress. There are many methods/styles of mixing audio, but when it comes to mixing surround sound, Fisher states: “There are no rules when mixing music in surround!” (2005, p. 88)

4. 3-DIMENTIONAL MOTION PICTURE (IN ANAGLYPH)

The recognition of depth comes from the fact of having 2 eyes, both eyes are looking at the same object, but from a different angle. For reproducing the human vision two identical video recording devices side-by-side with the distance of the human eye are needed. Of course, the distance varies for each individual but the calculated average should be approximately 65 mm to achieve the most realistic effect. Human vision does not only consist of having two eyes, there is also the most advance calculator or computer involved that is know to humankind - the brain.

"The principle behind stereography is relatively simple: it replicates human vision. With 20-20 vision, each of the human eyes sees a different image. Our eyes then converge at a certain point – which provides us with our sense of depth and three-dimensionality." (Atkinson, 2011, p. 141)

3D vision is very individual and it depends on the person watching and the system that is being used. In the last few decades the computing power has improved the evolving of 3D technologies; it has been made available for home applications, but it hasn't gained wide-spread popularity. "3D display technologies can be divided into four main systems: passive, active, virtual reality and anaglyph methods."(Dhaou et al., 2019, p. 1) The passive system uses special hardware (projector, screen, polarized lenses, etc) and polarized glasses, the active system uses a shutter-based system, the virtual-reality usually uses special viewing gear(glasses/occulars and headphones). "Finally, the anaglyph technique consists of two superimposed images of complementary colors describing the same view seen from a few offset points where the left image is in red and the right one is in cyan (green+blue). This technique is easier to apply and provides similar results to those obtained from other passive 3D

methods. Anaglyph techniques are the cheapest way to make the 3D visual experience achievable without the need for special hardware, but only colored glasses." (ibid)

4.1 Origin of stereoscopy, the predecessor of 3D vision

The basic principles of 3D vision have also evolved during several periods marked by significant technological developments. The first theories actually go way back to the time between 300 - 400 BC. "(...) Euclid stated in his manuscript *Optics* that depth perception is "to receive in each eye the simultaneous impression of two different images of the same subject".(*3D Video from Capture to Diffusion*, 2013, p. 11)

4.2 Stereoscopy

The principle of a stereoscopic device is to devide both eyes to see separately the same environment, but both pictures have been taken from a different angle, to create a sense of depth. The first known stereoscope (see figure 3.1) was introcued in the 19th century by Charles Wheatstone. "(...)Wheatstone's "stereoscope" (...), where two reversed images are reflected by two mirrors at an angle of 90 degrees (...)." (*3D Video from Capture to Diffusion*, 2013, p. 11) In the same era, others were also working on the reproduction of 3D images and improving the devices. Most known names to have been working on improving the stereoscope were Sir David Brewster and Oliver Wendell Holmes. For the spatial/3D effect to take place, a stereoscopy was used during the early years before and after the invention of photography (1840 -

Figure 3.1 Viewer Stereo 3D Film Glasses (Source: McMahan, G. 2004, pixbay.com)

Louis-Jacques-Mandé Daguerre in Paris, France and **William Henry Fox Talbot** in London, England).

In the beginning 20th century, the mass use of photograpyh and the development of motion picture escalated. “In 1915, the Astor Theater in New York held the first public projection of a short stereoscopic film entitled *Jim, The Penman.“*(*3D Video from Capture to Diffusion*, 2013, p. 11,12)

While shooting sterescopic motion there are many things to consider in the field of physics, cameras, lenses and camera alignment/rig. It is important to avoid close objects, that should “pop“ out of the screen, “touching“ the edges of the screen, this problem is referred to as *window violation*. There are also disparities that can be fatal for the 3D footage. The only disparity that is used to adjust the 3D effect is the horizontal disparity and this is called *parallax*. When slight problems occur with the camera rig like a small horizontal tilt, this can be fixed in post production, but this will also cause the video material to be a little bit zoomed in, and this is for the purpose of hiding the edges, that will give away the rotational fix.

> “The *interocular* separation (or interpupulary distance) technically refers to the distance between the centers of the human eyes. This distance is typically accepted to be an average of 65mm (roughly 2.5 inches) for a male adult. /…/ *Interaxial* separation is the distance between the centers of two camera lenses (specifically the entrance pupils.) The human interocular separation is an important constant stereographers use to make calculations for interaxial separation.” (Dashwood, 2010, p. 1)

Limitations for shooting 3D footage is mostly concerning blocking (distance the object is from the camera) and depth (distance between objects, foreground and background). The distance discrepancies are related to equipment that is being used. Each corresponding camera and lens will have optimal distances for creating spacious vision.

> "Binocular Vision and Parallax are the primary visual tools animals use to perceive depth at close range. The wider an animal's eyes are apart (its *interocular* distance) the deeper its binocular depth perception or "depth range." At greater distances we start to use monocular depth cues like perspective, relative size, occlusion, shadows and relation to horizon to perceive how far away objects are from us. Of course it would be difficult to look at double images all day so instead our eyes naturally angle in towards the object of interest to make it a single image. This is called convergence.(…) What never happens to your eyes in the natural world is *divergence*, which would mean that your eyes would angle outward." (Dashwood, 2010, p. 2)

4.3 Anaglyph method

> "The word *anaglyph* is from the Greek *ανα* = again, *γλμφη* = sculpture. In the classic method, used for monochrome stereo images, the left view in blue (or green) is superimposed on the same image with the right view in red. When viewed through spectacles of corresponding colors but reversed, the three-dimensional effect is perceived." (Dubois, 2001, p. 1661)

The benefits of anaglyph lie in the fact that it is inexpensive to produce and view. It is watchable on any screen and it provides the sense of seeing certain objects in 3D by stereoscopy. While it provides an inexpensive solution, it has many disadvantages including coloring and ghosting.

> "Anaglyph generation techniques are not too easy. (…) However, this technique is not efficient and suffers from three deficiencies which are the distortion of color, the retinal rivalry, and the effect of ghosting. In fact, the last one is a popular and major problem for most 3D displaying methods, especially for anaglyph 3D images. Indeed, it is due to the imperfection of the light-wavelengths filtering. As a consequence, undesired images leak through the red-cyan glasses, so they will be mixed with the required ones. The color distortion makes colors seen through red-cyan glasses and those of the original scene dissimilar while Retinal rivalry happens in

case similar objects in both scenes do not have the same colors for the left and right eyes. It is caused by the difference in the colored object brightness.

Therefore, it prevents concentration and causes visual fatigue and discomfort for long viewing. Despite these drawbacks, the anaglyph images are generally considered as the simplest and most uncomplicated methods for 3D representation.”(Dhaou et al., 2019, p. 2)

5. PREPARATION AND PRODUCTION

The idea of this project is to provide 3D audio-video content for the viewer to experience without expensive 3D equipment. The 3 components that are needed for viewing and listening are:

1. A screen (TV, laptop etc.)
2. headphones (or a 5.1 sound system
3. a selfmade paper/cardboard red-cyan glasses.

5.1 Realization

The band Puuluup was willing to participate in the process of audio recording and video shooting, the idea is to mix together visual content will be shot in the streets and the surroundings of Viljandi. The audio recording will take place in the video-studio of Cultural Academy of Tartu University and the band will be set up the way it would be also possible to film the recording-performance. The main idea for audio will be to record as many separate channels as possible for post production.

5.2 Audio

5.2.1 The audio equipment

For recording the band the mixing desk Presonus 32r will be used, mainly for the purpose of providing monitoring options for the band and through the desk it is possible to record a multitrack session via USB. The software for recording the multitrack session will the the freeware Presonus Capture. 48kHz will be used for the sample rate to provide as much syncronization as possible with the video files.

For recording the ambience in the nature and in Viljandi a portable 6-channel recorder will be used – Zoom H6 black edition. The recordings will be also in the sample rate of 48kHz for the recordings of these tracks will take place simultaneously.

Microphones for the vocals will be Telefunken M80 - a Supercardioid dynamic microphone and its frequency range, by specifications given in the manual, should be 50Hz – 18kHz, which is more than enough for vocals. The band will also use a AKG D7 that runs through an A-B pedal (a two-way selector between 2 channels for outputting on of them), which is connected to the effect pedals.

The ambience of the videoshooting will be recorded with the Sennheiser AmbeO microphone, which will be decoded from the 1-st Order Ambisonics A-Format to 2-nd Order Ambisonics B format in the mixing process. For the musical performance the same microphone will be used in the middle of the band, for creating the opurtunity of spaciousness in the music mixing phase.

10 D.I. boxes by BSS are also being used for splitting up audio during the music recording. These devices are being used for the precaution measures of grounding problems and for separation.

The live preformance of the band is reproduced with 4 channels:

1. Ramo Teder Vocals
2. Marko Veisson Vocals
3. Left channel of the looperboard Boss RC300 output (Talharpa and looping vocals)
4. Right channel of the looperboard Boss RC300 output (Talharpa and looping vocals)

The looperboard outputs create a stereo channel, which is the sum of all the different effects used during their performance.

For creating spacial audio, as many sound inputs as possible should be split before they enter the inputs of the looperboard. For separation of instruments the BSS AR-133 Active D.I. Box will be used. The input list for recording will be:

1. Ramo Teder Vocals
2. Marko Veisson Vocals
3. Ramo Teder Talharpa – split audio with BSS AR-133
4. Marko Veisson Talharpa – split audio with BSS AR-133
5. Ramo Teder Talharpa L after the loop of effects BSS AR-133
6. Ramo Teder Talharpa L after the loop of effects BSS AR-133
7. Marko Veisson Talharpa L after the loop of effects BSS AR-133
8. Marko Veisson Talharpa R after the loop of effects BSS AR-133
9. Looping microphone – split audio with BSS AR-133
10. “Kick-drum“ aka piezo microphone – split audio with BSS AR-133
11. Left channel of the looperboard – BSS AR-133
12. Right channel of the looper board – Bss AR-133
13. One condencer microphone infront of Ramo, low stand
14. One condencer microphone infront of Marko, low stand
15. Sennheiser Ambeo 1 channel

16. Sennheiser Ambeo 2 channel
17. Sennheiser Ambeo 3 channel
18. Sennheiser Ambeo 4 channel
19. Room microphone L
20. Room microphone R

Support for recording the audio in the studio and free-field recording will be by Michael Anthony Gugliotti. Most of the equipment needed for recording will be provided by Mulgikontsert OÜ and the BSS AR-133 D.I. Boxes will be provided by Estonian Traditional Music Centre.

5.2.2 Mixing and editing the audio

For mixing and editing the audio softwares Reaper and Pro Tools are used. For the surround effect to work on a 5.1 system, the main mix will be channel based – different sounds and noises will be panned manually into surround sound channels. For the Ambience to work in "stereo" or more precisely in binaural, first the A-format Ambisonics must be routed accordingly to a 4 channel bus then decoded to a B-Format Ambix , which after that will be processed down to a two channel track.

Since the main aim is still on the music, the surrounding noises are still there, but they are not highlighted and one could say, they are more behind volume-wise.

5.3 Video

The bands name Puuluup indicates to the word "puu" which means "a tree" or "wood" and they refer to their music during live-performances as of being about "trees and love"/"trees and food"/"food and sports"/etc. , which is the base for the idea for visuals – nature. The scripts for the songs have been created using phrases from the songs and imaginationational meaning of the lyrics which are, often described as, in multilingual language.

Locations of the video recording:

- Mudajärv, Viljandi
- the city of Viljandi.

5.3.1 The video equipment

For 3D video production will be used 2 exactly the same GoPro Hero 7 Black editions side-by-side using a dual flash bracket tripod double base mount on a Ronin S gimbal (a gimbal is a motorized device which is ment to support smooth camera movement in the hands of a trained camera operator). Both of the cameras will be equiped with external power bank for shooting in Linear-mode in the resolution of 4K and 50 frames per second(fps). The external power bank is suplied for the reason that in this mode the typical battery life is about 80 minutes.Testing the equipment has shown that it won't last more than around 40 minutes, during this time the camera starts to overheat, which means shooting precisely according to the script is needed in order to get all the material and breaks must be taken for the equipment to cool down.

As for the video side, the video-studio features a full-size green-screen backround for the video shoot, it is a necessary feature for adding the performers to the visuals that have been shot with a 3D camera rig.

Video support for operating cameras and filming the footage will be provided by Ako Lehemets and Jürgen Volmer.

5.3.2 Video editing and rendering

The raw video of the Gopro 7 Blacks will be converted firstly with the help of the program Gopro Studio, which actually has been discontinued since 2019. The software acutally provides better opurtunities for 3D conversion compared to the newer version Gopro Quik. The video material will

be edited and fine-tuned with the program Sony Vegas Pro (as this program has already integrated Stereoscopic 3D effects) and Adobe Premerie Pro for final touch.

The main problem with editing 3D footage, that has been shot with a “selfmade“-rig, is that one of the cameras needs to be aligned and tilted accordingly to mach the other camera. For these purposes Gopro Studio and Sony Vegas Pro have the means to do it quickly and without much effort.

A lot of footage will be shot in the Video studio, which means using a green-screen and chromakey in the post-production. The aim of this footage will be integration of the band into the nature and street shoots.

RESULTS

The aim of this project was to produce a short musical film that consisted of 3 different songs using the repertoire of the band Puuluup. While making the film some problems occured due to lack of technical equipment and unpredictable weather.

The recording of the audio in the video studio of Viljandi Cultural Academy's Vilma building went as planned. The AmbeO microphone was placed in the middle of the band in an upright position and this gave an extra option for creating space in the music. The recording of the ambience with the AmbeO microphone turned out to be more difficult than expected, because the recording was not static. Recreating the ambience that is seen in the videos was possible to a reasonable extent only. For the most realistic recording of the ambience in the forrest with the movement of the cameras and microphone would have required a silent "railway" through out the whole path that was visible in the video.

During the first day of the video shoot it was raining all day, which made it more difficult to achieve the desired result. The rain drops turned out to be a bigger problem than excpected in post-production. Most of these specific problems, like water drops on one camera, may benefit for the means of playing tricks on the viewers brain processing. It simulates the same problem, when someone is wearing glasses during a light rain and one of the glasses gets some water drops on it and the vision of the one eye gets blurry at some spots, but not fully. The solution for this problem is to keep the cameras out of the rain, because there are no plug-ins for correcting water droplets on a camera, as water bends light.

The second problem that occured was with shooting underwater. There was a lack of light, although the sun was shining directly on the water. The distance between the objects was to small and distance of the lenses was too wide for near-field shooting. These limitations made it impossible to fully correct the 3D horizontal offset. Since the average distance between the two eyes is 65 mm, the only available rig, that was delivered on time for the shooting, was a camera bridge with fixed distance of 69 mm, which is close enough for distant shooting. Under water the problem occured when the fish were too close to the cameras and it was almost impossible to correct the 3D stereoscopic effect to the point, where it was satisfying. One of the possibilities would have been to use one camera upside down, so that the lenses would be as close as possible, but the physical equipment for that was missing. For underwater shooting, a solution for recording the ambience was missing, which means that this part of the video is without ambience sounds.

Third problem occured when recording the ambience during the video shoot – since the camera and microphone were too close to each other but not close enough to the objects in the video, some of the recordings had to be re-recorded for the purpose of being able to use the Ambisonic ambience at all. Two recording sources being too close to each other will result in having footsteps of people in the recording that are not actually visible in the video. Typically video production uses a boom with a shotgun (for directionality) microphone, but in this case we used the AmbeO on the boom. The AmbeO microphone captured the sounds 360 degrees and that meant the captured audio had unwanted noises.

While watching the final outcome, it was made clear, that depending on the screen size and its color settings, the experience varies drastically. Adjustment of the screen settings and finding the right distance from the screen will improve the experience. To reiterate, each individual has to find their own sweet spot. The footage that has been created for Anaglyph 3D can also be converted to a side-by-side (active, passive and VR) 3D, which would get rid of the ghosting and coloring problems, but other technical challenges will arise.

During all the experimental 3D shooting, it was made clear that there is a certain distance that is needed for the optimal effect. Anything closer or further will not be seen as spacious. After trying a

couple of different „identical“ camera setups, the fact that nothing is identical in this world was confirmed. Using 2 „identical“ GoPro 7 Black cameras with the same setting, it still didn’t create exactly the same footage, which is vital for 3D content. Shooting on a side-by-side rig with a greenscreen proved not to be very effective using the method of Anaglyph 3D. The spaciousness is created together with distance of solid objects in the backround.

The video was shot in 2.4K 50fps instead of 4K 50fps, because 2.4K gives enough space to play with the 3D camera alignment and later render it in 1080p. Also the lower resolution for shooting video consumes less battery and space, which is important when being on a tight schedule and it need less processing power in the editing process.

For 3D horizontal offset correction differences between different takes, fading the videos proved to be the least painful transition for the eyes. Using jump cuts was confusing and disorienting, every cut was followed by a different 3D correction. The lightnes of the sceneries varied, because every video was shot on a different day with different weather. A fixed color balance, brightness, contrast and saturation adjustments could not be created and as a result all the 3 videos differ coloring-wise. The first video material was shot underwater in muddy waters, although the sun was shining straight on the pond, it was not enough for creating sharpness for the objects in the water. In the second video, it was a rainy day and there was no sun, so the whole scenery was abit dull and dark. The last video actually was blessed with a perfect weather for shooting in the forest, there was enough light and it was almost wind free, which allowed for great capturing conditions of audio and video.

CONCLUSION

The goal of this project was to create 3D audio-visual content using different surround sound and side-by-side camera techniques. The realization of this project showed that it is possible to create 3D audio and video, but the selected methods also have limitations. Finding the solution for these problems has been in progress for a long time and the 3D has gone way beyond anaglyph with the technical advancements.

The binaural and 5.1 decoding of the multitrack and AmbeO microphone provided accurate spaciousness that aided in the overall imaging of the audio. For adding extra virtual space, binaural panning of the room microphones helped to bring the sound more to the back of the listener.

Due to the constraints of COVID-19 the anaglyph method was the only method for other people to view the content at home without having to rent or buy expensive 3D equipment. This method also allowed for flexibility in schedule and meetings with the project members.

The project succeeded in creating spaciousness both visually and sonically, however some flaws and artificial disturbances occured. The binaural and anaglyph are methods are individual, since the visual and auditory perception varies from person to person. It is almost impossible to create an identical experience for all viewers.

BIBLIOGRAPHY

3D video from capture to diffusion. (2013). ISTE Ltd/John Wiley and Sons Inc.

Ambisonics Explained: A Guide for Sound Engineers | Waves. (n.d.). Waves.Com. Retrieved March 23, 2021, from https://www.waves.com/ambisonics-explained-guide-for-sound-engineers

Arteaga, D. (2015). *Introduction to Ambisonics*.

Atkinson, S. (2011). Stereoscopic-3D storytelling—Rethinking the conventions, grammar and aesthetics of a new medium. *Journal of Media Practice*, *12*(2), 139–156. https://doi.org/10.1386/jmpr.12.2.139_1

Ballou, G. (Ed.). (2008). *Handbook for sound engineers* (4th ed). Focal Press.

Corbett, I. (2021). *Mic it! Microphones, microphone techniques, and their impact on the final mix* (Second edition). Routledge.

Dashwood, T. (2010). *A Beginner's Guide to Shooting Stereoscopic 3D*.

Dhaou, D., Ben Jabra, S., & Zagrouba, E. (2019). A Review on Anaglyph 3D Image and Video Watermarking. *3D Research*, *10*(2), 13. https://doi.org/10.1007/s13319-019-0223-1

Dolby Atmos for the Home Theater. (2016). Retrieved March 20, 2021, from https://professional.dolby.com/siteassets/tv/home/dolby-atmos/dolby-atmos-for-home-theater.pdf

Dolby. (2021). Dolby Atmos Speaker Setup Retrieved March 20, 2021, from https://www.dolby.com/about/support/guide/dolby-atmos-speaker-setup/

Dubois, E. (2001). A projection method to generate anaglyph stereo images. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, *3*, 1661–1664 vol.3. https://doi.org/10.1109/ICASSP.2001.941256

Fisher, J. P. (2005). *Instant Surround Sound*. Taylor & Francis.

Hallum, R. (2017). *Stereophony-a series of perspectives—What exactly IS stereo?* https://doi.org/10.13140/RG.2.2.10854.11844

Holman, T. (2008). *Surround sound: Up and running* (2nd ed). Elsevier/Focal Press.

Kim, M. Y., & Peters, N. G. (2020). *Flexible Rendering of Audio Data* (Patent No. 20200105282). https://www.freepatentsonline.com/y2020/0105282.html

McMahan, G. (2004). Viewer Stereo 3D Glasses https://pixabay.com/photos/viewer-stereo-3d-film-glasses-1069128/ (10.03.2021)

Neumann. (2021). Dummy Head KU-100 https://en-de.neumann.com/ku-100 (10.03.2021)

Postrel, S. R. (1990). Competing Networks and Proprietary Standards: The Case of Quadraphonic Sound. *The Journal of Industrial Economics*, *39*(2), 169. https://doi.org/10.2307/2098492

Roginska, A., & Geluso, P. (Eds.). (2018). *Immersive sound: The art and science of binaural and multi-channel audio*. Routledge, Taylor & Francis Group.

Rumsey, F. (2001). *Spatial audio*. Focal Press.

Sanftmann, H., & Weiskopf, D. (2011). Anaglyph Stereo Without Ghosting. *Computer Graphics Forum*, *30*(4), 1251–1259. https://doi.org/10.1111/j.1467-8659.2011.01984.x

Tsakostas, C., Floros, A., & Deliyiannis, Y. (2007). *Real-time Spatial Mixing Using Binaural Processing*. 5.

SUMMARY

The aim of this project was to create 3D audio-visual content with the music of Puuluup. For spatial audio elements, the sound of the band and ambience, was recorded with an Ambisonic microphone by Sennheiser, the Ambeo VR. The audio was mixed down to binaural and up to 5.1 surround. The visual content was recorded with two GoPro 7 Black cameras that were rigged in a side-by-side method to reproduce human vision.

Experimenting with the binarual, 5.1 and anaglyph methods for 3D were utilized during the undertaking of this project. Although binaural reproduction has its benefits and disadvantages, it is suitable for creating surround sound for headphones. The 5.1 system provides an even better surround image than binaural, but this lacks the perception of height. One must also consider that the channel order for exporting 5.1 surround for reproduction purposes.

The anaglyph method proved to be a working and inexpensive solution for creating 3D video. Although this method has many limitations, it has gained a lot of popularity in the field of photography and other artistic mediums. The limitations are being praised yet they are the very same deficencies that make this technology seem outdated.

Through the making of this project many technical and stylistic measures were tested. Despite the limitations and restrictions, the outcome was satisfactory in reproducing audio-visual 3D content in an easily viewable way.

Non-exclusive licence to reproduce thesis and make thesis public

I, Viljar Rosin,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Puuluup Immersive,

supervised by Christoph Felix Schulz, Janar Paeglis and Jürgen Volmer.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Viljar Rosin

17/05/2021