TAMPERE UNIVERSITY
OF APPLIED SCIENCES

SOUND PRODUCTION FOR 360 VIDEOS

In a Live Music Performance Case Study

Mark Malyshev

Bachelor's thesis
May 2018
Degree Programme in Media and Arts
Music Production

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Degree programme in Media and Arts
Music Production

MARK MALYSHEV
Sound Production for 360 Videos

Bachelor's thesis 34 pages, appendices 7 pages
May 2018

The purpose of this thesis is to overview and analyze the production process of creating 360 videos.

This thesis suggests why sound should be recorded in a precise way to give viewers a deeper experience in the VR universe. It will briefly go through the basics of sound recording, explaining related theoretical information about psychoacoustics and ambisonics that would also be needed in practical applications.

Music and sound post production is a fundamental part of audio media content. These days Virtual Reality is getting more and more popular, calling on audiences to enjoy a new experience; however, in most cases production teams pay more attention to the video component, even though sound has an equivalent role in making viewers believe in what they see.

On the practical side, this thesis will explain the work process of sound production for VR and 360 videos by going through the thesis project, which is live recordings of a rock band made for Tampere University of Applied Sciences, as an example of the workflow and work process.

Key words: vr, sound productions, spatial audio

CONTENTS

GLOSSARY

Ambisonic	Full-sphere surround sound technique
Psychoacoustics	Studies about sound perception.
Binaural	Recording and playback sound technique creating a 3D effect. It is recorded and played back with two channels, that is why sometimes it is confused with “stereo”.
VR	Virtual reality is a technology that uses headsets or mobile phones with players that allows the viewer to “look around” inside the video.
FB360	Software designed for creating spatial audio for 360 videos.
Spatial audio	Sound that exploits sound localization.
Inside the band	The surround sound approach that appeared in early 1970’s attempted to deliver a new perspective and experience to the listener, that would place instruments around but not in front of the listener. (Holman 2000, 14.)
EQ	Process of reducing or boosting certain audio frequencies to change the sound perception.

1 INTRODUCTION

Virtual Reality or VR technology was in development for many years, and in many forms; but only nowadays has it finally acquired the necessary technical support to become a mass product, and to be widespread in media all over the world.

These days technologies provide production companies with online players and services with wide accessibility. The production of hardware has improved, and that has lowered prices making the technology more accessible to consumers. The percentage of VR users is growing and that requires development in content and quality.

One of the main VR branches is entertainment. People are willing and ready to spend enormous amounts of money to be entertained, and VR is a new way to satisfy consumer demand. However, it has one additional feature compared with normal media entertainment - it gives the consumer an experience of new dimensions and opportunities. However, this innovation involves many technical issues. Those issues are mainly due to modern production quality standards being very high. The quality of video, sound and content that is industrial standard has nearly reached the top and many production companies started to look for new ways to surprise their consumers and compete with other companies.

This competition stimulates development of new technical devices and new possibilities based on old ideas and dreams that in the beginning of 20th century sounded like futuristic insanity. However, the execution quality should be very similar to the usual 2D video and stereo sound, because that is what the consumer is accustomed to and that is what they expect to get. It is not enough just to make video and audio move around the viewer. The quality of those two components should be as good as in usual media - or better, despite the fact that the video consists of several cameras stitched together and interactive audio that utilises a transforming phase and frequency spectrum to follow the listener’s head movements.

All these factors of how media react at VR as a product or extra service shows that subject should be studied. Since there is still not much information online about certain cases gone from the beginning to the end it has to be researched by practical cases.

In other words, it is necessary to understand the standards of the audio industry before starting to work on VR pre-production. In this thesis, I will mention the main fundamental aspects of audio production that should be taken into consideration.

While going through the entire production process, in the case that I will describe in this thesis, I have faced the fact that it is a new field on which very little educational literature is available. In the case of usual live production, it is easy to find many books and educational videos that explain the production process, using the particular equipment that you own, in detail.

The practical case on which this thesis is based is a shot and recorded concert of a rock band called “69 Eyes” playing at Tampere Talo (The 69 Eyes – Jetfighter Plane 2017). In addition to the practical case I will compare different mixing approaches.

Thus, in this production I will use live music recordings, post-production mixing, sound design and spatializing audio for first-order ambisonics as a final format. The major online video players accept mp4 video files with B-format ambisonics, as do offline players for Samsung gear VR, Oculus Rift and HTC Vive. A knowledge of accepted file formats and appropriate software that can be used is very important in increasing the production’s speed and quality.

2 EXPERT INTERVIEWS

2.1 Tim Gedemer

Tim Gedemer is an American sound designer, known for his work with the True Crime and Need for Speed games. He has been involved with sound effects for media professionally for 23 years. (http://www.vgmpf.com, 2016.) Tim is also owner and VR audio director of Source sound (SourceSound 2017).

2.2 Jukka Holm

Jukka Holm is a senior audio engineer in Nokia Tech. He is mainly involved with VR productions projects. (Holm, 2017). Jukka has experience in working as a production manager for VR/360 videos (LinkedIn 2017).

2.3 Antti Immonen

Antti Immonen is technical development producer in YLE production and design department, in Mediapolis (Tampere). He has been in YLE since 1985 and has been doing numbers of projects in audio area in this company. Antti was a technical manager in RSO performing Strauss symphony project, and was investigating with a team the practical way how to make a multi camera recordings from the stage and using separate recorders. (Immonen, 2018.)

3 VIRTUAL REALITY SOUND

3.1 Standard technique

One of the most popular and used online VR players is YouTube. YouTube is an online service to watch and share originally-created videos. (YouTube 2017.)

The VR video uploaded to YouTube should be in mp4 format and should include mono, stereo or 4 channel multitrack recordings. The last one also should be supported with metadata. (YouTube 2017.)

Multitrack audio file should be 4 channels, ACN channel ordering, SN3D normalisation with first order ambisonics (FOA) (YouTube 2017).

According to the player request we should be able to make certain audio format. There are several possibilities to get necessary audio format.

3.2 Ambisonics of the first order

The source of an ambisonic signal may be an ambisonic microphone or it may be artificially panned mono signal, split into the correct B-format components and placed in a position around the listener by adjusting the ratios between the signals. (Rumsey, 2001, 112.)

The first order ambisonics is currently widely used in 360 sound recordings and post-production.

3.2.1 A-format

A-format is used for recording spatial audio by ambisonics microphone. One of the examples is Ambeo microphone by Sennheiser. A-format consists of four sub-cardioid capsules orientated as shown in figure 1.

FIGURE 1: The A-format microphone Ambeo, developed by Sennheiser, uses four capsules. (© Sennheiser 2017)

A-format is used for recording spherical sound and usually it is converted to B-format in the post-production process (Sennheiser 2017).

3.2.2 B-format

B-format consists of four signals that between them represent the pressure and velocity components of the sound field in any direction as shown in figure 2.

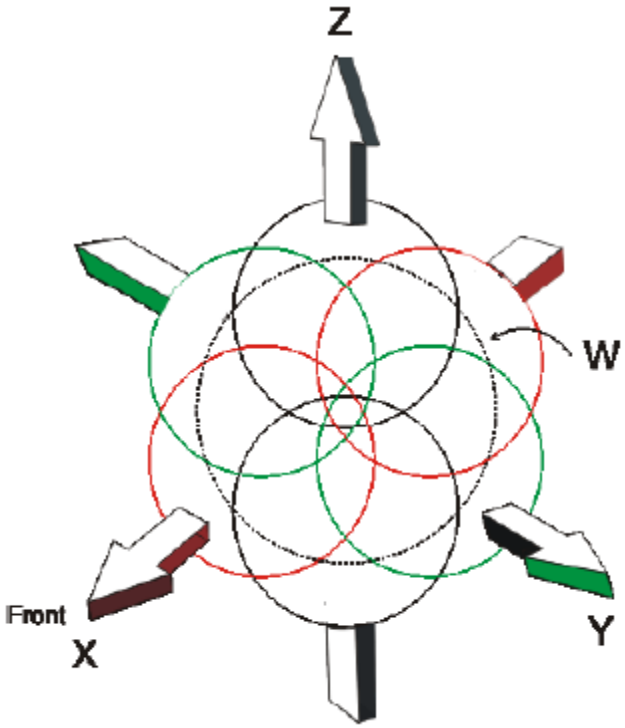

FIGURE 2: The B-format components W, X, Y and Z in Ambisonics presented by 3 figure-eight and one omni patterns. (Brown, 2010)

It can be seen that there is a similarity between the sum and difference format of two channel stereo, since B-format is made up of three orthogonal figure-eight components (X, Y and Z), and an omni component (W). All directions in horizontal plane are possible whilst Z is required for height information. X is equivalent to a forward-facing figure-eight (equivalent to M in MS stereo), Y is equivalent to a sideways-facing figure-eight (equivalent to S in MS stereo). The X, Y and Z components have a frontal, sideways or upwards gain of +3dB with relation to the W signal (0 dB) in order to achieve roughly similar energy responses for sources in different positions. (Rumsey, 2001, 113.)

While working in a DAW, B-format is widely used to create 360 sound. It requires 4 channel tracks, thus in Pro Tools it is necessary to use an HD system designed to provide surround sound. However, DAWs such as Reaper allow work with 4 channel tracks without an HD system or any extra hardware. (A REAPER User Guide v 5.70, 2017, 58**)**

3.2.3 Normalization

For successful reconstruction of the sound field, it is important to agree on a normalization method for the spherical harmonic components. SN3D format was approved for that purpose. SN3D is easier to use when dealing with audio data: clipping of the higher order signals can be avoided, because the peak amplitude of single point sources will never exceed the level of the 1'st order signal; in practice, this turns out to be more relevant than the N3D normalization for statistical diffuse fields. (Nachbar, Zotter, Deleflie, Sontacchi, 2011.)

This normalization is usually included in software dedicated to 360 mixing, so it is not necessary to add it separately in the production process.

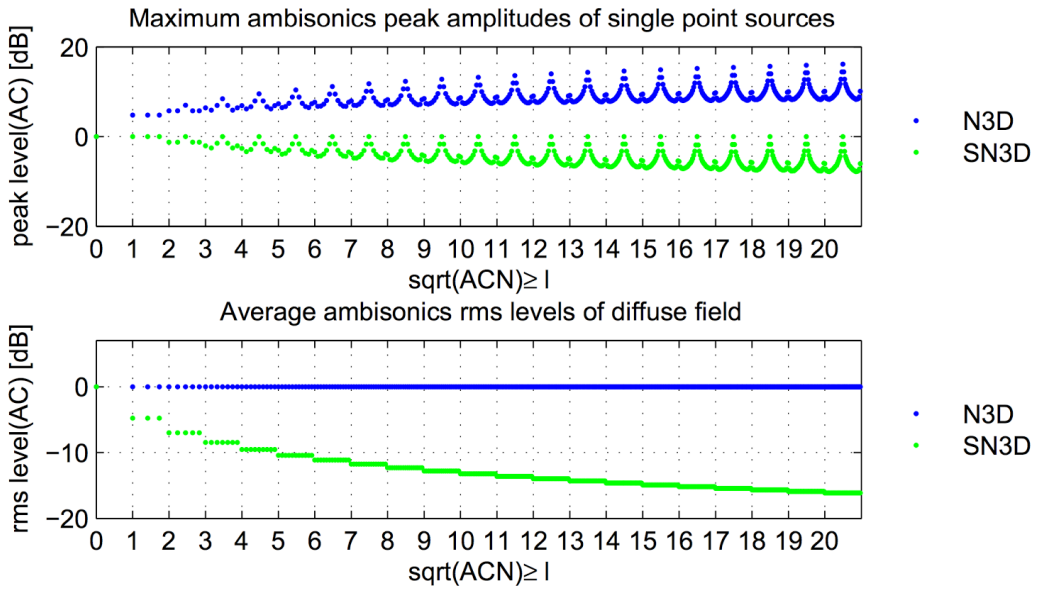

FIGURE 3: Peak and RMS levels by N3D and SN3D normalization with ACN channel ordering (Chapman, Ritsch, Musil, Zmölnig, Pomberger, Zotter, Sontacchi, 2009)

3.2.4 Channel ordering

ACN is an ambisonic channel ordering system that was proposed and documented in 2008 for future ambisonics files. (The Ambisonics Association, 2008.) However, it is not the only channel order that is used in the industry - for example Wwise import files should have FuMa channel ordering and maxN normalisation. (© Audiokinetic 2017.)

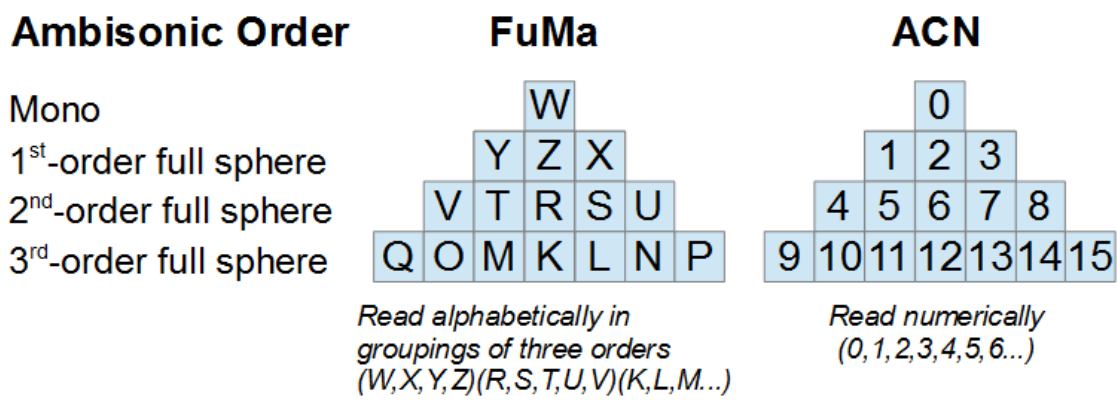

FIGURE 4: Ambisonics primary component ordering. (© Audiokinetic 2017)

It is important to follow the same channel order as is in the software used. For example, Ambisonics Tool Kit (ATK) uses a FuMa channel order and it is not suitable for YouTube videos. In this case, the channel order should be changed before the audio is merged with the video. Other software uses an ACN channel order thus the production process is made easier.

3.2.5 Metadata

Metadata injector is a tool for manipulating spatial media (spherical video and spatial audio) metadata in MP4 and MOV files. It can be used to inject spatial media metadata into a file or validate metadata in an existing file. (Spatial Metadata Injector 2017) In other words metadata is needed for the player to work with information about the audio's position in space. Without implemented metadata, the player will reproduce a still sound image. Implementation could be done by ffmpeg or tools using ffmpeg, for example FB360 Encoder.

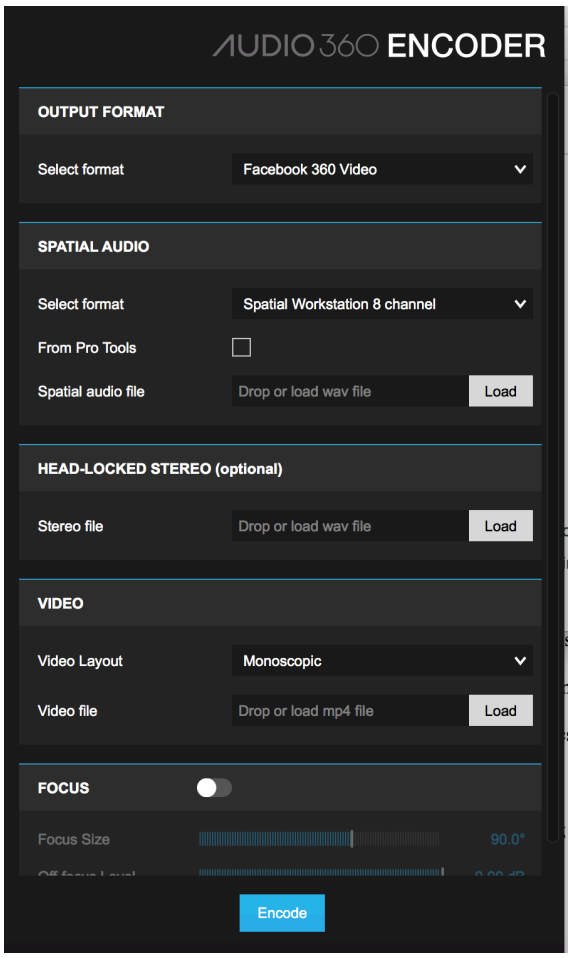

FIGURE 5: Encoder interface (Malyshev, 2017)

There are many options available for combining sound to make ambisonics, but as no standardised option existed, it was impossible for them to become mass consumer products. Once the most logical and preferred option was found, it was ready for the next step. The new chapter of ambisonics and VR started when this format was chosen by Google.

In my work, I would like to focus on the particular format that was described in the beginning of this chapter. The format in question is YouTube request, as I found it to be fully suitable for the purpose of this project.

4 VR PRODUCTION IN PRACTICE

There is a variety of ways to achieve a final product; however, I would like to focus on one possible way and will briefly outline some alternative options. The production process I will describe will be based on live video and audio recordings of a rock band.

The event took place in Tampere Hall in April 2017. At pre-production stage, we analysed the stage and gear in order to have a clearer vision for the recording process. With the video crew we planned camera position and moving range. We decided on the optimal place for the sound recording crew and commutation plan.

For the shooting of the video, we used a Nokia OZO camera.

FIGURE 6: OZO camera (©Nokia, 2017)

This device has 8 built-in microphones that create spatial sound (©Nokia 2017), however, it was decided that we would use multitrack recordings from the sound desk for this project. The sound pressure next to the stage during live shows is strong, and it would be impossible to have any kind of control over the sound in post-production.

4.1 Psychoacoustics

> Psychoacoustics relates the measurable physical properties of waves, like amplitude and frequency, to the perception of sound and subjective phenomena, like loudness and pitch. It is psychology part of sound. (Farnell 2010, 77.)

Despite the fact that it is subjective and more related to feelings and perception, it is very important to achieve the necessary sensation for the viewer; in other words, to build an immersive effect.

Fundamental elements of localisation of sound sources in space are interaural time difference (ITD) and interaural intensity difference (IID).

ITD is the time difference between the arrival of the same sound at each ear. If it arrives at the right ear before left it is very likely that source is located at the left side. The most effective frequencies for ITD are below 700Hz and it stops having an effect over 1.5kHz. This effect is related to the length of the wave and size of our heads. (Farnell, 2010, 79–80.)

IID is the difference in intensity - in other words, loudness. Simply put, if the sound is louder in the right ear than in the left, it is very likely that the source is located in the right field. (Farnell, 2010, 79–80.) However, it is necessary to mention that in daily life we barely stay with our head still at all, and this causes changes to ITD and IID. It is subconsciously used to localize a sound source. In the mixing process, it is very important to be able to play 360 video properly synced with audio. This is important in order to make the correct panning adjustments and also to determine the size of the object.

4.2 Basics of sound

Recording VR footage requires that the engineer knows the fundamental theory behind sound and recording. Firstly, I would like to mention a couple of things about the basics of sound.

We receive and transform sound signals as so-called sound-pressure waves. Solid objects always vibrate in reaction to environmental changes. Objects such as speakers are designed to vibrate, squeezing air molecules into compressed area next to them and this causes differences in air pressure. This process is called compression. As the vibrating object moves inward from its normal resting state, an area with a lower-than-normal atmospheric pressure will be created, in a process called rarefaction. This process is visually presented as wave. (Huber, Runstein, 2013.)

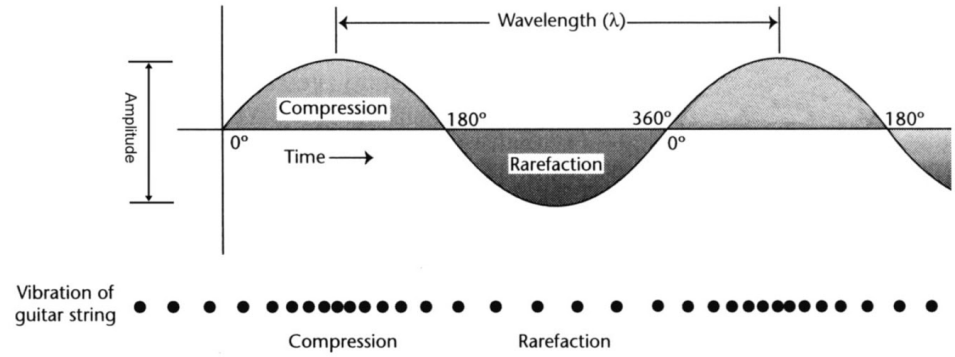

FIGURE 7: Visualisation of a guitar string soundwave (https://theproaudiofiles.com/understanding-sound-what-is-sound-part-1/)

Every waveform is represented through several characteristics that make one waveform distinct from another.

- Amplitude
- Frequency
- Wavelength
- Harmonic content
- Envelope

Amplitude is the distance above or below the centreline of a waveform (such as a pure sine wave). It represents the level of the signal. The bigger

the distance between top or bottom from the centreline, the more intense the pressure is. Amplitude can be measured in several ways. For example, the measurement of maximum positive or negative signal level of a wave is called peak amplitude level (or peak level). The total measurement of the positive and negative peak signal levels is called peak-to-peak value. Average level of a waveform over time is called the root-mean-square (rms). (Huber, Runstein, 2013, 45–46.)

Frequency is the rate at which vibration repeats within a cycle of positive and negative amplitude back to centreline of waveform. The number of cycles that occur within a second is measured in hertz (Hz). Frequency influence pitch of a sound, more repeats occur the higher pitch is. (Huber, Runstein, 2013, 47.)

Wavelength of a waveform is the physical distance between the beginning and the end of a cycle (frequently represented by the Greek letter lambda, ƛ). In relation with pitch the shorter wavelength is, more cycles would occur within a second, the higher pitch would be. (Huber, Runstein, 2013, 48.)

Harmonic. Up to this point sound has mainly been described using examples of sine waves to simplify the definition, however in a real-life environment we usually do not hear pure sine waves, and all the instruments have voicing – a combination of frequencies that makes the sound unique, otherwise all the instruments would sound similar. Harmonics higher than the fundamental pitch are called overtones; those that are lower are called sub harmonics (Huber, Runstein, 2013, 54). Harmonics and overtones are important in describing sound character, that is why understanding the basics of them is critical. To describe what actual sound we hear we have to take into consideration all the aspects of the sound source or instrument, such as the type of material the instrument is made of; the size of it; the vibration of the string, reed, and membrane; vibrations of all materials and how they are assembled; and so on. The natural waveform that we hear is a combination of all those aspects combined with acoustics by a law of physics (Gibson, 2011).

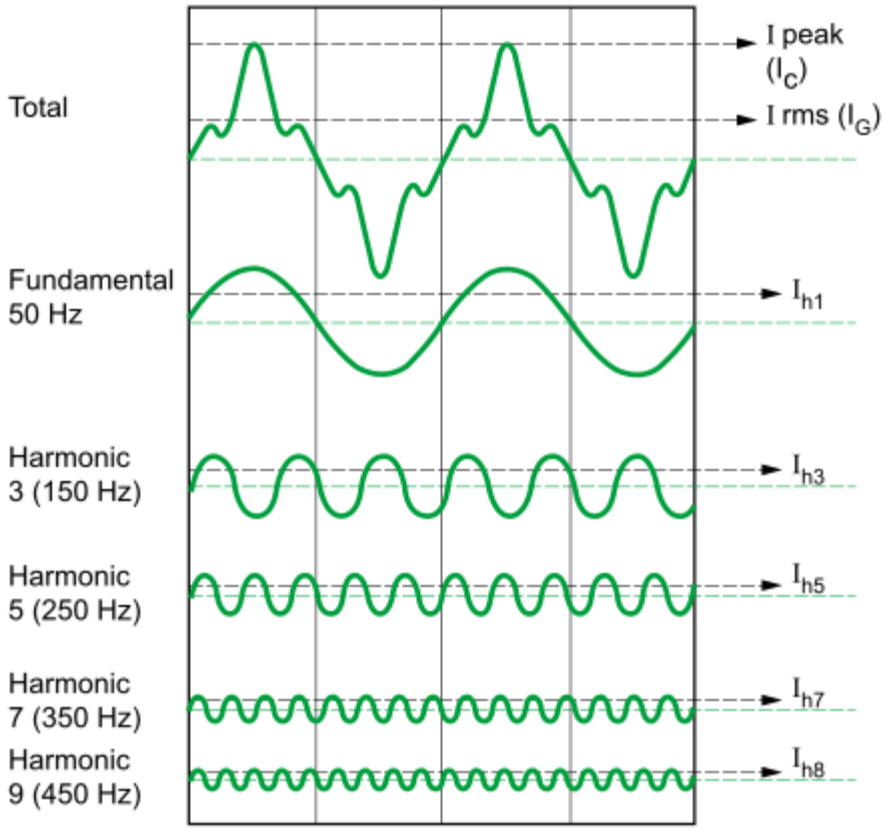

FIGURE 8: Representation of the physics of harmonics and their overall influence on the final wave; peak and rms (http://philotone.com/)

Envelope is an initial action, development and diminishing of a waveform over a period of time. It includes four main phases of the envelope of a waveform: attack, decay, sustain and release. Attack is the phase of the envelope when sound is initiated. The amplitude rises from the centreline to peak level. There are instruments and sound sources with fast attack such as clap, snare drum, bell; and slow attack such as bowed violin or reversed crash cymbal sound. (Gibson, 2011, 51–52.)

Decay. When the peak attack is reached, the energy might decrease quickly thereafter until it reaches constant amplitude. The phase of decreasing level after attack is called decay. (Gibson, 2011, 51–52.)

Sustain. Once the sound has levelled from the attack and the decay, the period that the sound is still generating from the sound source is called sustain. As long as the source continues, the waveform is sustaining. The sustain phase can remain at a constant amplitude, increase or decrease in amplitude. (Gibson, 2011, 51–52.)

Release. Once the source stops generating the sound, the envelope enters the release phase. The best example of release phase is reverberation. (Gibson, 2011, 51–52.) However, the release phase is not just reverberation, once a musician stops playing cello the strings and body of the instrument are still resonating and good musicians with good instruments make release an important part of their performance. (Gibson, 2011, 51–52.)

Our brain collects many details and information from each phase of envelope, sound pressure, pitch, and harmonies, to create an overall picture of the sound.

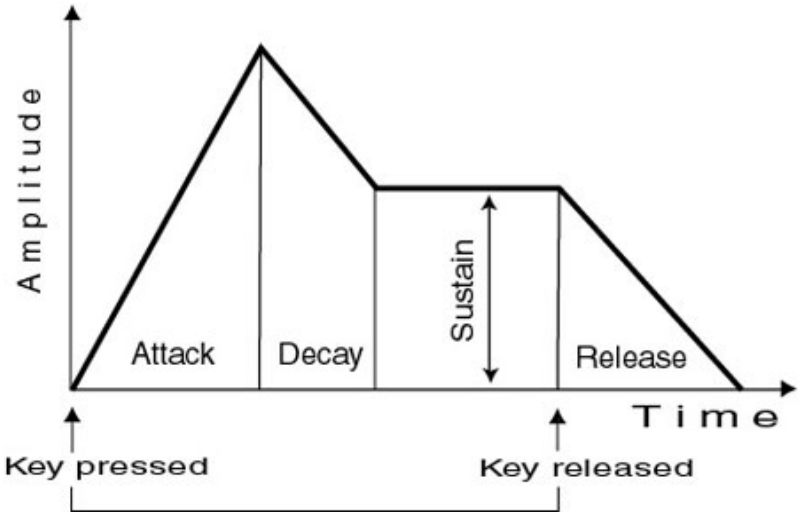

FIGURE 9: Sound Envelope ADSR represented by keyboard instrument or synthesizer (http://making-music.com/)

4.3 Recording

Recording and shooting are a fundamental part of audio and video production. Most of the distance and perspective information we perceive through visual content and good synchronization with audio. (Mendonça, Mandelli & Pulkki, 2016.) That is why we have spent a good amount of time in pre-production for this case study. In 360 video production camera positions are crucial, that is why we accurately planned them with video crew to make sure that the distance to musicians would be optimal, and that there would be enough lights and good overall view. However, video was not the only aspect we were paying attention to; in spatial audio we planned to make automation for every

sound source and make them move with musicians, that means that there will not be accurate static balance as in stereo or static surround sound mix. That means that we should choose a position for main camera in a way that it would provide the viewer with balanced mix, and several extra cameras to show an “inside the band” perspective. Before the actual show we contacted the band members to show them how the final result could look, to give them an idea of what kind of movements they could use on stage to create a beautiful performance. The difference is that in the end the viewer would feel as though he or she were on the stage, and that the artists were performing just for them - thus looking at the camera is not forbidden as in 2D shooting. (Holm, 2017.)

Another option is if camera is mounted on the camera crane in other words the Jib, or mounted to the camera dolly – a wheeler cart that is used for shoots. That allows changes in camera position or the making of motion cameras. (Holm, 2017.)

The reason why paying attention to the camera position and range of possibilities of motion is important, is that it helps to plan post production for audio and calculate the time needed for it. Representing perspective and distance naturally in post-production is also getting easier. With multiple sound sources, their size and distance become a crucial aspect, however in the case of making musical footage there must be a balance between making a “realistic” representation and a “musical” approach.

Balancing between those two aspects is subjective thus communication with customer becomes critical in finalizing a project. There are two main approaches: making a realistic spatial representation of camera position and perspective of audio in space, and the other one is to follow composer’s or mixing engineer’s idea of balance of instruments. The first approach will give a good immersive effect to the sound but it will be disbalanced musically. Representing camera position in the audio would make the nearest sound sources much bigger and louder than further ones. However, the second approach allows you to keep the artist’s image and sound, especially if the sound and quality is a fundamental part in artist’s performance. The immersiveness will not be as good as in the first option, because our expectation from what we see and what we hear in the final product will make dissonance. (Holm, 2017.)

Those are the main video and directing aspects that help to create an immersive experience for customers, however I would like to focus on sound production. Usually in our

productions we just need to record a multitrack session with clear healthy sound. However, there could be cases when we need to stream live, and then mixing should be done before sending files to the internet.

In the case study recordings, we were working with the aforementioned mentioned rock band, which consists of drums, bass, guitar, backing track and vocals. It is important to save some extra footage from recordings - additional information that could help with the mixing process.

Our main goal for those recordings was to capture clean, unprocessed sound with full dynamics and no distortion.

4.4 Mixing

Before the mixing process, it was important to agree on the song that was to be produced. We went through video and audio material and agreed on the song that would be the most interesting to present in VR video.

Before mixing to ambisonics it is necessary to prepare balanced mono and stereo stems. The number of stems should be equal amount of sound sources we see in the image. When stereo mixing is done and approved with the customer it could be used as a reference track for the 360 version. Such elements as drums, keys, grand piano, backing track may be prepared as stereo files to be able to reproduce their natural size and width. It is also important to record and use ambience microphones to capture the natural depth and acoustic color of the place. Using several pairs of ambience microphones would allow you to make the overall space sound smooth and natural, even so, when mixing in 360, extra artificial reverberation is usually required.

However, there is one fundamental difference between making sound for usual video as opposed to VR. Standard DAWs like Pro Tools (with the exception of the 12th version of Pro Tools HD), Logic or Reaper do not have a video player that would work with 360 video. Thus, it is necessary to find a VR player that could be also synced with the DAW used. Fortunately, the company “Two Big Ears” has designed the plugin ‘FB360

Spatial Audio Workstation' that allows users to sync their VR videos using a video player working in slave mode. Using that plugin, it becomes possible to use a standard DAW to create audio files with 1st and 2nd order ambisonics.

Once the DAW was synced with the VR player it was necessary to synchronize the video with sounds. It was done in the classic manual way: however, in further projects we have used sync generators to keep multiple cameras and audio in sync. Mixing has become a combination of classical mixing and VR approach. FB360 works with panning in space influencing EQ response, and delay between left and right channels once the ambisonics sound is converted from ambisonics to binaural playback engine. It also works with depth of soundfield; however, in my opinion one FB360 plugin is not enough to create a natural feeling. For that purpose, a combination of EQ and compression could be used for those sounds that always stay in the distance. Creating a sense of space is a crucial aspect in 360 mixing. (FB 360 Spatial workstation user guide, 2017.)

One remarkable feature of 360 mixing is that binaural audio is based on HRTF (Head-related Transfer Function) that is calculated in anechoic chambers, and in the post-production process we are reproducing a live environment that consists of reflections. Thus, it is important to recreate the room and its behaviour. FB360 has implemented a basic room model to improve sound source localization. (FB 360 Spatial workstation user guide, 2017.)

5 MIXING APPROACHES

5.1 Classical orchestral mixing

Most of cases that we have had were related with a multi-camera shootings and post production. Since there is no strict standard in mixing music for Ambisonics but there are classical approaches to mix live music, our goal was to experiment and compare different approaches in mixing. As an example of more classical approach in ambisonics mixing I have interviewed Antti Immonen, who was working on Helsinki Radio Symphony Orchestra's performance of Strauss symphony. (Immonen, 2018.)

The approach that was chosen for the final version of that concert is closer to classical ways of mixing orchestra and based on Decca tree. Decca tree is foundation in orchestra recording and mixing process, represented by five microphones place over conductor. (Huber, Runstein, 2013, 66)

That choice was determined by several factors. First of all, the concert hall is shared between two orchestras with different orchestras and that makes preparations more difficult. The audio crew was not able to use their own microphones and was limited with 64 microphones from the hall. Microphones had fixed position and most of them were rigged from the ceiling thus they covered wide area that did not allow to emphasize proximity and details of small group of instruments. (Immonen, 2018)

The 5.1 mix that was made for the broadcast was recorded and used as a reference to ambisonics mix. After mixing and listening session in Nokia's Spherical control room, which has spherical immersive control system with lots of Genelec speakers it was clear that spot microphones in FOA do not give precise specialization. Thus, working with higher orders ambisonics was sensible. On that moment, the new version of Pro Tools HD 12.8.2 was released and it allowed to work 2nd and 3rd order. (Immonen, 2018.)

Since the time for production was short the combination of 5.1 mix inverted into ambisonics in combination with spot microphones for percussion and lower range of string section was reasonable. To emphasize proximity and camera position, OZO microphones were mixed into the main mix. (Immonen, 2018.)

That approach allows to keep balanced mix from Decca tree and add spatial details to the final mix. However, it sacrifices the real representation of sound sources position in space and that could be crucial for VR video. Although Decca tree is converted to ambisonics, the sound and video perspective fits only camera which is placed next to a conductor. Inside the band perspective. Figure 10 illustrates how audio placement of objects in space could be different from video.

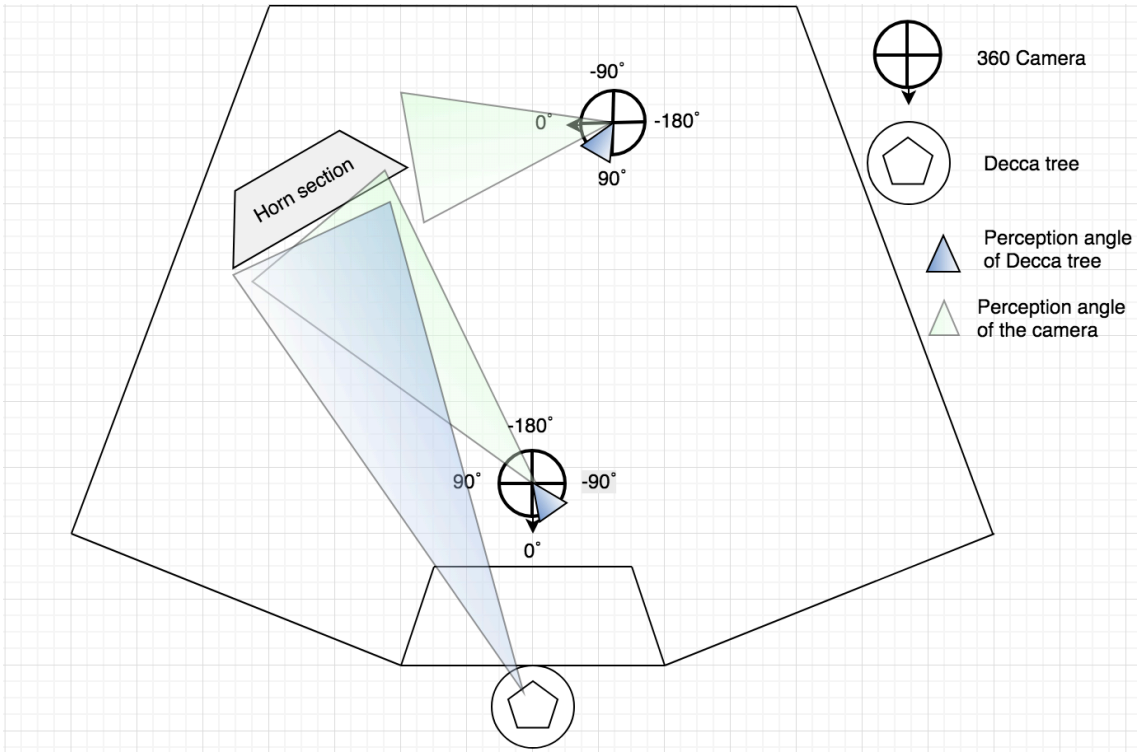

FIGURE 10: Example of panning stereo into spatial audio. (Malyshev, 2018)

5.2 Alternative orchestral mixing

In a way, it is possible to compare usage of Decca tree for orchestra and standards in 5.1 mix for other styles of music. In both approaches direct sound is placed in front 90˚ – 110˚ of surround. 360 and VR music live videos are recorded using ambisonics microphones or using static stereo mix.

However, if there is possibility to setup spot microphones on stage close to instruments, then it is possible to create a mix out of spot microphones avoiding using Decca tree. Balancing spot microphones with artificial reverbs would allow to create smooth transitions. In that approach, it is necessary to follow the same logic and create ambisonics mix that would be as close as possible to 5.1 reference. Then after panning audio chan-

nels to the sound sources position it is necessary to add some level adjustments to emphasize depth and proximity. That approach also allows to have precise control on every channel reverb sends.

Every mono channel converted to 2nd order ambisonics consisted of 9 tracks. If an orchestra consist of 40 pieces that would be 360 channels to process and reverbs in addition. Thus, it is reasonable to prepare mix for 5.1 or quad and then convert submix channel into ambisonics. That would allow to save computer power and keep the phase clean.

6 OVERVIEW

Here I would like to sum up the audio production process in a bulleted list with short comments according to the practical case. Here, the process will describe recording a band without A-format microphones for one camera. In case of multi camera production it is necessary to create mix for each camera and then create project for editing between cameras.

• Audio recording.

• Making stereo mix. It is necessary to confirm it with customer and have it as a reference.

• Preparing mono and stereo stems. The number of stems should fit the amount of sound sources in the picture. The stems were prepared in Pro Tools.

• Create separate project for 360 mixing.

• Import 360 video to the DAW. There are many combinations in playing video with DAW – (in this case Reaper and FB360 as video player are used)

• Import audio stems to the DAW. After they are imported mono or stereo should be converted into ambisonics (FB360).

• Sync video with audio. In the case that audio and video are synced with the same clock it is easier, otherwise synchronization should be done manually.

• Make panning placement for static sound sources. As the starting point, it is necessary to position all the sounds at their sources.

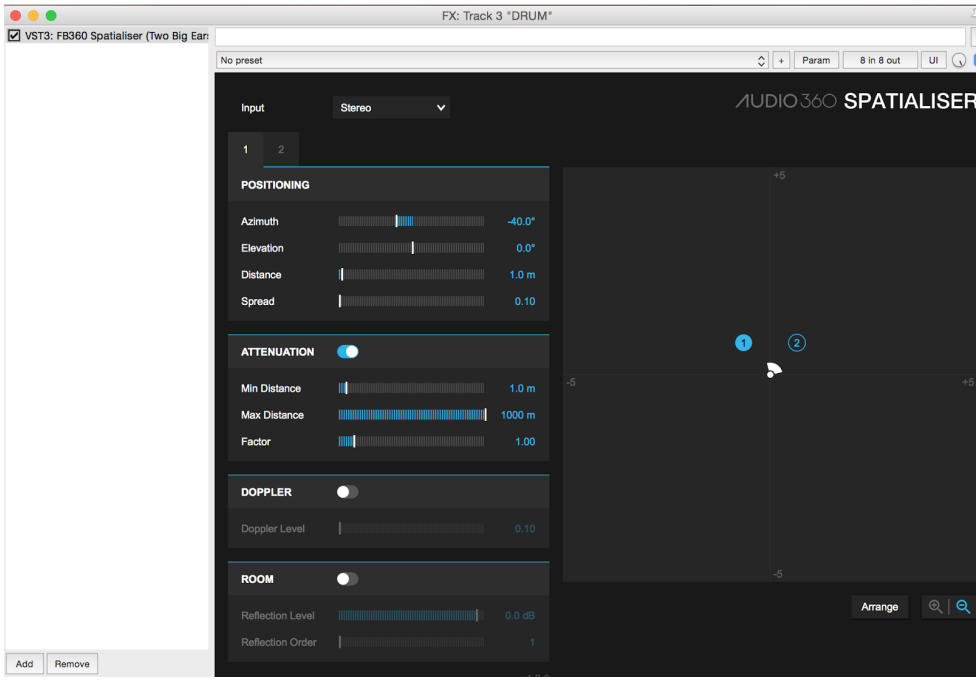

FIGURE 11: Example of panning stereo into spatial audio. (Malyshev, 2017)

• Create acoustics and space. In most of cases VR video needs extra reverberation processing.

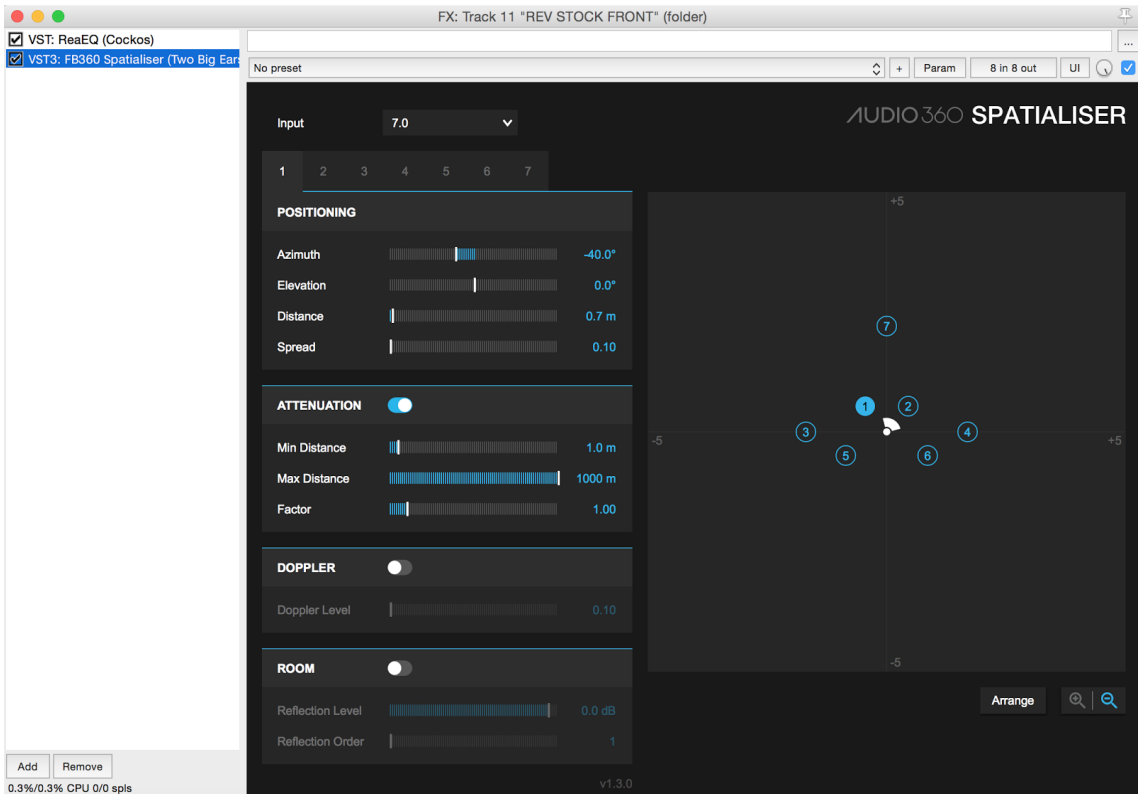

FIGURE 12: Pattern of artificial spatial reverberation. (Malyshev, 2017)

• Make automation. If sound sources move it is necessary to make automation so sound follows them.

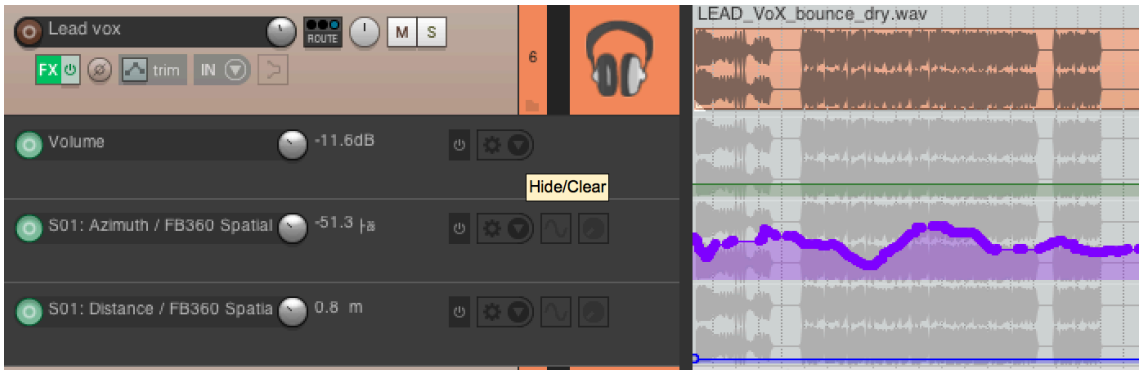

FIGURE 13: Reaper Azimuth automation. (Malyshev, 2017)

• Exporting audio. For YouTube 4-channel track, or 9-channel (2nd order ambisonics) if it is used for several players.

• Merging sound with the video using Encoder. It is important to know how the video will be used. Unfortunately, players still do not have one standard thus the video should be adapted to different players.

• Ready to upload.

7 DISCUSSION

VR video and sound production is a process that is still under development, thus it is constantly changing and it is hard to find any standard solutions and standards in general. That makes production unstable and variable. Development depends on customer requests and customers have not shown much of an interest yet for several reasons. First of all, the amount of content currently being produced is small and production includes big costs as well as being a complicated process. Production companies do not want to spend money on "risky" projects that would barely pay off. The second reason is quality. Although some production companies spend money to provide customers with high quality content, the industry is still not entirely ready for it. For example, YouTube video quality is decent only if viewers have powerful enough devices to play videos and the internet connection is nearly perfect. Otherwise, the quality decreases and consumers are not interested in watching it. That causes bad feedback and comments from viewers about quality of content and also decreases their interest in it. These are two of the most fundamental questions of VR: accessibility and quality. Obviously, none of those factors can be solved until customers show their interest - in summary; all these driving factors influence each other.

Modern production of 2D video and sound is of such high quality that the average viewer no longer accepts anything less than a top-quality product. It is not only about technical quality: directors are using their skills based on all the possible tools to create amazing material. The 360 world is still developing the visual aspect: effects, the quality of stitching and quality of playback, and it will take time to bring them to the same level with 2D production. As we have already seen, most obstacles are related to video production and not sound. However, it is important to mention that those two elements are very important to each other. Our perception glues them together so close that it is difficult to focus on quality of sound while watching a blurry image that is constantly dropping frames because of poor internet connectivity. Likewise, the immersive effect of a high definition video fades away if the sound is static and of low quality.

However, our topic is sound, and in sound production for 360 videos it is important to know and understand all the basics of classic stereo and surround sound mixing techniques and be able to apply them in practice. The VR world is based on surround sound

represented by binaural audio, thus basic knowledge of psychoacoustics is valuable as well. All small elements like 360 panning, automation, facility acoustics, size of objects and sound are important and create an immersive effect if they are in balance with the image. Otherwise the viewer notes dissonance between image and sound. In most cases they are not able to identify the specific elements that in their opinion cause that dissonance, but the immersive feeling is already lost.

Whereas 360 mixing does not have any certain standards yet, it is easy to experiment with mixing techniques and be creative. With the speech, it is straight forward panning and syncing, but sound design and music mixing require more artistic solutions to create. In case of music shooting and recording priority is always on the side of the video and ambisonics microphone should be as close to the camera as possible, but that would not give you perfect position for sound recording.

With VR getting to the music industry the need for different types of projects has increased. All of them require audio post production and using only ambisonics microphones is a limited option. Thus, it is important to study and practice multitrack mixing and converting the mix to ambisonics.

Outstanding branch of audio production for VR is producing music. That process requires from audio producer all the fundamental knowledge about the nature of the sound, recordings and VR audio features to be able to combine those together. Ambisonics sound has certain features of representing positioning and proximity that are not acceptable in classic stereo music. In other words, matching the ambisonics mix with stereo reference is one of the most important parts in mixing process. First of all, after stems preparation it is necessary to write precise automation to follow to sound sources in the video. There are several ways to achieve decent proximity effect: first of all, some plugins like FB360 allow user write information about distance to the object into automation curve, however, it also could be achieved by slight volume change and decreasing early reflections frequencies and attack. Using second option usually slow down the mixing process, but also gives more accurate controls for the engineer.

Another important aspect is creating the space that would support immersive feeling. There are also several options to solve that task. First is to use an ambisonics microphone next to the camera that would create audio field but in case of shooting rock con-

cert from the stage one microphone on stage is not an option. For that case, an ambience pair or pairs of microphones and could be used. However, those microphones cover narrow area and deaf spots could appear. To avoid that it is good to create artificial space with reverbs. Such plugins as FB360 or Audio Ease suggest users ambisonics reverbs and rooms to create natural space. However, that could be achieved by manual combining of several reverbs. The last one allows engineer to have precise control on every single spot of the reverberation whether it should be longer pre-delay or brightness of reflection from each direction. That technique allows an engineer to have controls on how much every single instrument or sound could be send to any of the reverbs depending on needs of the mix. That could be well used in mixing orchestra in ambisonics when the camera is inside the orchestra. It allows to keep the mix closer to original balance, create height and width of the space, and not depend on one reverb.

Structured workflow could help to decrease many phase issues. In case of multitrack mixing it is necessary to separate motion objects and static ones. For example, orchestra is static and it is reasonable to prepare mix for quad first and then convert it to ambisonics instead of converting every single close microphone into 9 channels ambisonics track.

Since every camera position require separate mix to represent correct position reasonable would be to mix every each one of them and then edit in separate edit session. But in actual edit version it is getting clear how transitions between elements work. Thus, in some cases to save time and make efficient mix it is necessary to create project with all of the audio elements presented on one surface. To achieve that it is necessary to make session for quad mixing first, place every element at its own place and then use secondary bus to ambisonics plugin that uses quad as source. Then it is necessary to create the space using reverbs and separate those reverbs according to their position is space. For every position in space reverbs ratio is different and level of sends should agree with visual perception. Those reverbs should be sent to master send of the camera. Those steps should be repeated for each camera and by using cameras master faders it is easy to correct balance and create smooth transitions between positions.

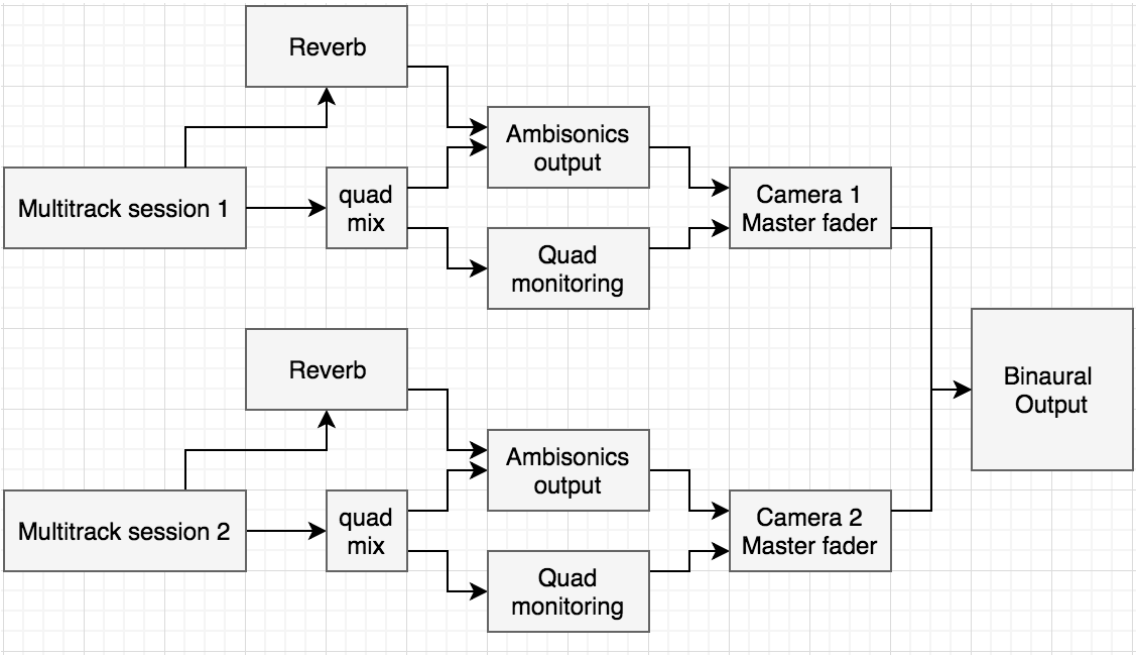

FIGURE 14: Scheme of routing for multi-camera mixing session. (Malyshev, 2018)

Mixing sound for 360 videos is about precise panning and syncing automation envelops with the video material and using many audio channels (since every 2nd order ambisonics track includes at least 9 channels), thus structure of the project itself is crucial. It is important to keep project simple and structured so video playback would be precise and work with minimal latency.

REFERENCES

Audiokinetic. Using Ambisonics. Read 23.11.2017.
https://www.audiokinetic.com/library/2016.1.0_5775/?source=Help&id=using_ambisonics

Brown B. 2010. 3D Microphony. Read 23.11.2017.
https://www.prosoundtraining.com/2010/03/11/3d-microphony/

Chapman M., Ritsch W., Musil T., Zmölnig I., Pomberger H., Zotter, F. Sontacchi A., 2009. A Standard For Interchange of Ambisonic Signal Sets Including a file standard with metadata. Read 23.11.2017.

Farnell, A. 2010. Designing Sound. The MIT Press.

FB 360 Spatial workstation user guide, Read 17.01.2018
https://facebookincubator.github.io/facebook-360-spatial-workstation/Documentation/SpatialWorkstation/SpatialWorkstation.html#what-did-you-download

Gedemer, T. 2016. Tim Gedemer: Sound for VR. Interviewer Glenn Kiser. Read 23.11.2017
https://soundcloud.com/soundworkscollection/sound-for-vr-with-tim-gedemer-of-source-sound

Gibson, B. 2011. Live Sound Operator's Handbook. 2nd Edition. Hal Leonard Books.

Holman, T. 2000. 5.1 Surround Sound Up and Running. Focal Press.

Huber, D., Runstein, R. 2013. Modern Recording Techniques. 8th Edition. Focal Press.

http://www.vgmpf.com/Wiki/index.php?title=Tim_Gedemer, 2016.

Jukka Holm, LinkedIn, 2017. Read 23.11.2017
https://www.linkedin.com/in/jukkaholm/

Nachbar C., Zotter F., Deleflie E., Sontacchi A., 2011. Ambix - A Suggested Ambisonics Format. Read 23.11.2017.
http://iem.kug.ac.at/fileadmin/media/iem/projects/2011/ambisonics11_nachbar_zotter_sontacchi_deleflie.pdf

Nokia, OZO, 2017. Read 23.11.2017.
https://ozo.nokia.com/en/products/create/ozo-virtual-reality-camera.html

Rumsey, F. 2001. Spatial Audio. Focal Press.

RSO and Hannu Lintu: Summary of Richard Strauss Alpine Symphony. Read 30.05.2018

https://www.youtube.com/watch?v=nwmO8AWnYCU

Sennheiser, Ambeo VR Mic, 2017. Read 23.11.2017.
https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic

Sourcesound, Read 30.05.2018
http://www.sourcesoundvr.com/about/

Spatial Metadata injector, Google, Read 30.05.2018
https://github.com/google/spatial-media/blob/master/spatialmedia/README.md

Tampere Philharmonics, Highlights of the Moomin museum grand opening gala concert. 360 music video. Read 30.05.2018
https://www.youtube.com/watch?time_continue=3&v=K5tW00CikH8

The 69 Eyes, “Jet Fighter Plane”. 360 music video. Read 30.05.2018
https://www.youtube.com/watch?v=cdipTX_kbvQ

The Ambisonics Association, 2008.Amibsonics Channels. Read 23.11.2017.
http://ambisonics.ch/standards/channels/

YouTube, Upload instructions and settings, Upload virtual reality videos, 2017. Read 23.11.2017.
https://support.google.com/youtube/answer/6316263?hl=en

YouTube, Upload instructions and settings, Upload virtual reality videos, 2017. Read 23.11.2017.
https://support.google.com/youtube/answer/6395969

YouTube, 2017. Read 23.11.2017.
https://www.youtube.com/yt/about/

APPENDICES

Appendix 1. Interview Jukka Holm 20.11.2017

I would like to ask about planning for 360 video production and how is it important to choose good position for the camera and how does it influence final result?

Well we have learned that the closer we can get to the performer the better it will look, especially with VR glasses. And if the performers are close to the camera the viewer actually has to move the video, he has to rotate it to all directions to see the performance and especially if they move it could improve the experience, so if the cameras are too far away and if you see all the performance all the time in the same spot then you really don't want to move the video, and then you kind of lost the idea of 360 video. So, if you are too far away from the performance it become kind of 2D video. It is also good idea to have of the cameras "inside the band" kind of, so if they all are in front of the stage then you see the audience on the one side of the camera and the band on the other side of the camera, and sometimes the audience is not so interesting to look at. It is always the compromise you have to discuss with the band members where you can place the camera. So, you can think about the cameras like the human head or as a person standing on the stage.

We have also used the jib several times and we were able to move the camera closer to the guitar player through the solo parts, for example. We have found it very useful.

So, would you say that it extend the amount of possibilities?

Yes. You can make slow movements, but you have to be careful that people could feel sick when they view with glasses, or you can just quickly move to the perfect spot. It is not possible to use the dolly because of space restrictions. For example, we are not allowed to use rails in our next concert because the place is sold out and whey will not have that much space in the front of the stage, and there will be photographers also. But when it is possible it looks very good.

How does it change sound production when you use those tools?

Well it makes more work for the sound engineer because when the dolly or the jib moves the engineer has to pan all the sound sources, but for the viewer it is good experience. As I said before if the stage is big we can change position for the dolly or the jib and leave it there if musicians are moving, it gives us more flexibility to capture the best part of the performance.

Would you say that video has more value that the audio in VR?

I guess that quiet many people would be happy with just the stereo sound, people are not used to mixing techniques that we are using because basically they did not exist one or two years ago. So, video is still more important, for example when we have done live streaming the sound was stereo and people were happy with that because the video looked so good.

Do you think that spatial audio has value in VR videos or would it be just fine to use stereo instead?

I think that spatial audio has big value and it increases the immersiveness but people have also been happy with the stereo version. Of course, we always want to do our best and we will do spatial audio whenever that is possible.

What are main steps in audio recordings?

It is always important to discuss with the sound engineers how can we record the sound because every time the consoles are different, they use analogue desks or MADI or DANTE. It could be anything and there are so many different types of cables and standards, so the step one is to discuss with the sound engineer what do they have and try to collect the necessary gear for that particular event. So far, we were pretty successful with that because we had careful preparations. Also, it depends if we are doing live streaming then we need a stereo mix then it gets trickier because we have to do stereo mix ourselves live and that could be problematic. But I would say that the most important part is to find the phone number of sound the sound engineer and contact him so there will not be any surprises.

Can you underline main steps of audio processing?

Usually we record several songs and then the band picks one and to choose it they want to hear the rough mix, they do not want to see the video, they trust us, but they want to hear the audio in case if they have made mistakes. Also, if the sound engineer of the band wants to check the mix we are making stereo version and send it to get approval, after that we typically start making the spatial mix based on the stereo mix. Then in the DAW we are using prepared stems and number of stems should be the same with the number of sound sources. We start with placing those stems to match the video and for that we need some kind of low quality version of 360 video so the audio engineer can work with them. With FB360 plugin we were able to play the video in the DAW and place the stems, convert them into first order ambisonics and make automation for them.

Also when we recordings we might want to have ambience pair of microphones or several to add them to the 360 mix and make it sounds natural but we also add artificial reverbs to make sounds better. We use artificial because we have controls on their settings. It helps to keep the feeling of musicality and the style. For example, if you compare classical recordings or rock concert, the settings are different.

What about synchronization? How is it done?

We have struggled a lot with it. First productions we have made synchronization manually in post-production process, but nowadays we are making synchronized video recordings. We are using Blackmagic design HyperDecks and there is one HyperDeck for each camera and they are synchronized with the same timecode and all the HyperDeks started recording at the same time. That makes post production process much easier.

How do you export the audio after you have finished mixing?

Typically, it is first order ambisonics because at the moment Facebook and YouTube support it so it is kind of defacto standard, I would say.

How do you merge audio with video?

To do that we use FB360 Encoder, that uses ffmpeg coding to do that, but you should know what are you making the video for. For example, YouTube or Facebook, or Nokia 360 player has different settings so you really have to know for which purpose the video will be used so audio works correct. For us it has been trial and error process.

Appendix 2. Interview Antti Immonen 18.05.2018

Hello, can we start with a small introduction?

My name is Antti Immonen and I am technical development producer in YLE production and design department, here in Mediapolis (Tampere). So I have been in YLE since 1985. I have been doing everything in audio area in this company that possibly could been done.

Also, you have done orchestra recordings and post production?

Yes, in YLE we have been doing orchestra recordings. Definitely, my experience is more in studio and live entertainment music shows, but I have been doing some operas and classical recordings, so it is not that unfamiliar to me.

And what is your opinion about VR and 360 videos, in general?

Since I wear eyeglasses I am not sure that current virtual reality glass technology is quite there. It is not that fascinating yet, but the glasses are developing all the time, they become lighter and quality of picture is coming better all the time.

Both of us have been working on recordings of Radio Symphony Orchestra of Helsinki that took place in may of 2017. Could you please tell something about it?

This project was part of the business Finland funded research project, we were researching practical ways to produce VR media in masses. What would be the workflow and what does it take to produce content in VR world. And we were researching the production from audio, video, graphics, story-telling and delivery points of view. I was a technical manager in that project, and was investigating with our team the practical way

how to make a multi camera recordings from the stage and using separate recorders. Everything about cabling and synchronising issues, so that all recordings should be done at the same time and the same speed. So that was the basics and secondary goal was how would sound behave with cuts and picture dramaturgy works with the multiple cameras within VR world, because there is no consensus around the world, should we cut it or not.

And you have been also mixing audio for the video?

Yes. And since the same stage shares two orchestras: HKO and RSO, thus the schedule was very strict and we could not place our own microphones there. We had to use hanged microphones from the hall. We have got 64 channels of audio and also the 5.1 mix that was done in the control room for the live broadcast. The live mix was very handy in the end as a reference and we also have partly used it in the final mix.
First I was using FOA (First Order Ambisonics) tools in Reaper (DAW) to create spherical sound field. Also I had a mixing and listening session in Nokia's spherical control room, which has spherical immersive control system with lots of Genelec speakers, and we immediately heard that it doesn't work the way we wanted. The FOA resolution is not enough to represent all necessary spatial elements. Although, the balance was good and sounds were good, it was hard to clearly point any instruments in that mix. Then we switched to ProTools HD when 12.8.2 version was just released. This version has 1st, 2nd and 3rd order ambisonics tracks in there and all necessary ambisonics tools. The second attempt was to use ProTools in combination with FB360 Spatial Audio Workstation tools for mixing and panning channels. Since the time was running out, we have decided to use 5.1 master mix which was placed into the spherical audio field and we added the first order ambisonics audio from OZO cameras to each position and also some spot microphones, mainly for percussion and contrabass area.
I was trying to use hanged spot microphones to make more precise sound for the inside-the-band camera positions, but since mic position was quite high, there was no use those microphones to represent close sound. There should be much lower and closer to the instruments to be used in VR Ambisonics mix.

What was the final ambisonics format? 1st order or higher?

Delivery format was Spatial Workstations 8 channel modified second order mix for Youtube

How does 1st order of decca tree mix is different from close microphone one?

Do you mean difference between FOA session (Reaper) and SOA session (Protools and Spatial Workstation)? Spatial resolution is poor when using Deccatree and FOA.
When using FOA with Ambeo or similar as an FOA tetrahedral main microphone, you have a possibility to use something like Harpex-X to enhance spatial resolution.
When using FOA, it's beneficial to move microphones closer to sound source to increase spatial pinpoint of instruments and signals.

How is height represented in ambisonics using 5.1 foundation? Did you use artificial reverbs in addition?

Yes, reverbs are good at this to create illusion of height. We did have an ambience microphone pair hanging above audience, which also was partially used for height immersion.

What reverbs were used and in what ratio?

We tried several reverbs from Valhalla and UAD (Lex 224 emulation) they are all very good, but several instances are needed to form a quad layer of reverbs. User experience is not the best possible.
Then in the end, I purchased Exponential Audio's Phoenix Surround reverb plugin. Phoenix has a 3D link feature, where you can link two instances of plugins, ie. horizontal 5.1 and quad height plugins. You can also crossfeed audio between instances. The sound from Phoenix reverb was very good also.

In first tests with FOA (using mainly spot microphones) how was the mixing process made for that?

In the Reaper FOA session, I created folders (=subgroups) for every instrumental sections (woodwinds, brass, VLN1, VLN2, VLA, VLC, CB, Perc…etc), folders were stereo subs. At folder channelstrip there was an ATK 3D panner to pan the stereo sub into

a spherical audio field (in its real position in an orchestra). These 3D panned stems formed an ”close mics”-subgroup in FOA format. This group was mixed then with spatialised 5.0 Decca, FOA Ozo-tracks, Ambience (height) etc… But as mentioned before, it was a mess, with no localisation or anything. Close mics were hanging too high to get some close orchestral section sound from them.

How was done audio edit between cameras?

Our director, Ilmari Huttu-Hiltunen was cutting picture with Premiere. I was getting the edited equirectangular 4k video from him which had to be downconverted to HD resolution for Protools video timeline.

I made an separate editing session, where there was only 5.1 mix (converted to SOA), OZO audios (rendered to FOA then upmixed to SOA) and Cb and percussion spotmics (with SOA panners).
Then I put the edited picture into video track, and synced the video and audio.

Since we were having only 8 minutes of material to be produced, I edited OZO and spotmic audios manually when nudging and detecting cuts from the video timeline, it was fast and easy. There were relatively few cuts in the picture. In real world some kind of an AAF workflow should be created, with automatic cut detection.

After audio splicing, I simply muted the regions, that were not part in the point of view video, ie. when cut to the OZO #3 (percussion), all other regions were muted except Ozo #3 audio and it’s spotmics. Then after that, I made some fade ins and outs if needed and finally there was some fader automation over everything.