EFFECTS OF LOSSY COMPRESSION CODECS ON THE PERCEPTION OF IMMERSIVE AUDIO IN VIRTUAL REALITY APPLICATIONS

Sabian Hibbs
University of Derby
100602673@unimail.derby.ac.uk

ABSTRACT

This research investigates if the use of Vorbis lossy compression on ambisonic audio that has been binauralized within Audiokinetic Wwise, and auditioned through Unreal Engine 4, will change the directional accuracy of the information encoded in the ambisonic audio with respect to the compression ratio. A listening test is then preformed and evaluated, drawing current conclusions from data gathered from the listening tests and presented in this report.

1. INTRODUCTION

The use of ambisonic audio in video games has not been widely adopted, the primary reason for this is a data storage issue, this is due to the multiple channels higher ambisonic audio creates when recording. Ambisonic audio at 3rd order $(19\ Channel - 24\ bit - 48Khz - 1152\ Kbps\ Bitrate)$, using the lossless PCM – Pulse Code Modulation format produces $2.736\ Megabytes\ (MB)$ of data per second. That is $164.16\ MB$ per minute of recording, and with ever growing need for more intricate sound systems in games, this can cause a problem with memory budgets. With modern gaming platforms such as the Microsoft XBOX ONE only able to store $< 250\ MB$ of audio files for any one project using the S.H.A.P.E Audio Engine [11]. This causes a situation where developers do not wish to sacrifice the memory budget for ambisonic audio. Due to the localization benefits of ambisonic audio, with respect to the audio being binauralized over headphones, can compression give some benefits to the storage volume issues inherited by ambisonic audio, without losing the localization of audio within. This report will evaluate and analyse the localization effects of ambisonic audio that has been compressed with the lossy codec Vorbis, specifically due to Vorbis' being the primary compression codec used in the game audio authoring tool Wwise [13].

2. VORBIS [.OGG]

Vorbis is an open source, non-patented lossy audio codec format that has its technology derived from a number of functions such as vector quantization and transformation within the frequency domain, also known as the modified discrete cosign transform ($MDCT$), all processed through a psychoacoustic model derived from the limitations on human hearing.

The specifications for mono audio at 48 kHz show that the maximum bitrate of encoding is 250 kbps. Wwise enables pre-sets for audio quality when encoding in Vorbis.

Preset	Quality	Nominal bitrate
		Xiph.Org Foundation Vorbis
	-2	not available
	-1	45 kbit/s
	0	64 kbit/s
	1	80 kbit/s
Auto Low	2	96 kbit/s
LOW	3	112 kbit/s
Auto Mid	4	128 kbit/s
MID	5	160 kbit/s
Auto High	6	192 kbit/s
HIGH	7	224 kbit/s
	8	256 kbit/s
	9	320 kbit/s
	10	500 kbit/s

Figure 1. Vorbis Bitrate to Quality ratings from (-2 < 0 < 10) with reference to auto detection quality (Auto) present in Wwise compression settings for Vorbis [10].

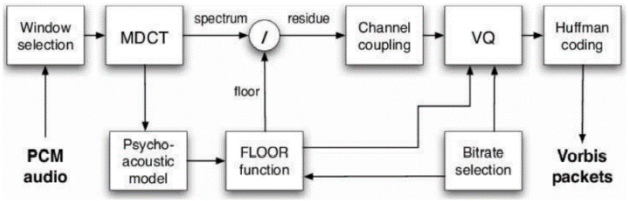

Figure 2. Proposed flow diagram outlining .Ogg Vorbis Encoder [1].

$MDCT$ is a transform based on the discrete cosign transform (DCT) but with the added lapped transform. The lapped function in signal processing is a class of linear discrete block transformations, where the basic function of the transformation overlaps the block boundaries, yet the number of coefficients overall resulting from a series of overlapping block transforms remains the same as if a non-overlapping block transform had been used [3].

Equation deriving the $MDCT = X_k$

$$X_k = \sum_{n=0}^{2N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right] \quad [3]$$

Where:

R denotes the set of real numbers, $2N$ are real numbers.

$$F: \boldsymbol{R}^{2N} \rightarrow \boldsymbol{R}^{\boldsymbol{N}}$$
$$X_{0,} \cdots, X_{2N+1}$$

2.1 Vorbis Psychoacoustic Model Coding

psychoacoustic model coding is a set of lossy compression codecs that try to remove data from audio that the human auditory system would either not notice, or know perceptually information was there to be removed. The human auditory system has a generalized frequency sensitivity of around $20\,Hz$ to $20\,kHz$, although this frequency band is not the same with every person, as humans are inherently different from one another. Suitable approaches can be made using this data for the removal of frequencies beyond the general model. Shown in Figure 3 is the psychoacoustic model used within Vorbis.

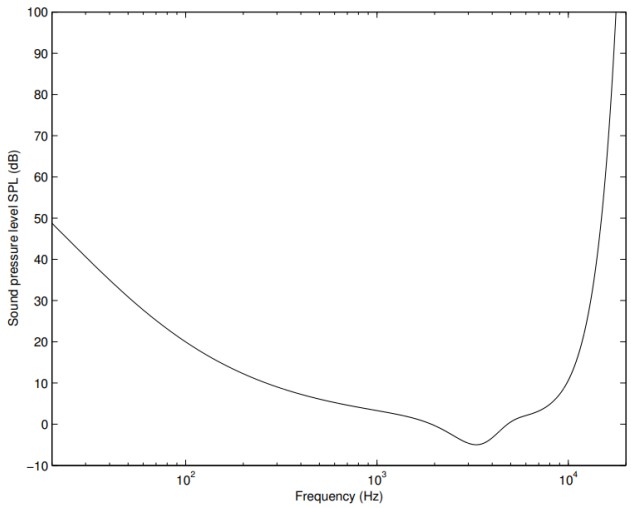

Figure 3. Absolute threshold characteristics of the human auditory system $T_q(f)$ with reference to Vorbis [2].

Figure 3 can be approximated using the following equation:

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5_e{}^{-0.6\left(\frac{f}{1000-3.3}\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \quad [2]$$

Using this information Vorbis identifies the bands of audio that do not conform in the psychoacoustic model shown in Figure 3. With the use of windowing in the frequency domain, the signals outside of these bands are filtered out.

2.2 Vorbis Critical Bandwidth

Critical bandwidth is the frequency range that subjectively can change at fast intervals. In short the human brain's ability to gather information sonically is limited to some degree, as some sonic transitions within this critical bandwidth are not perceived by the brain. Critical Band Rate or (CBR) can be used in tandem with the bark scale. Bark corresponds to the physical structure of the human auditory system, and can be used within this model as it has a general linearisation of the human ears auditory response that shows linear characteristics for low frequencies that alter slightly for a more logarithmic characteristic for higher frequencies. To convert from linear frequencies to CBR the following equation can be carried out [2].

$$z(f) = 13\arctan(0.00076f) + 3.5arctan\left[\left(\frac{f}{7500}\right)^2\right] \quad [2]$$

Where: z is the Bark and f is the frequency.

2.3 Vorbis Masking

Masking is a term that is defined in signal processing, as frequencies at a specific point in time being close to other perceived louder frequencies, thus masking the original sonic content. Vorbis uses a model based on both the psychoacoustic model and the critical bandwidth, to calculate frequencies that can be filtered.

3. Research Inspiration

Research that specifically tests auditory localization with ambisonic audio through a binauralized system shows that the higher the proposed order in ambisonic recording, increases the band of frequencies that are able to be accurately reproduced [5]. Further studies have shown that compression of surround sound audio ($5.1\,to\,7.1$) systems, with respect to Vorbis, do have an auditory effect on the encoded audio when testing on a number of people [1]. A number of studies have accurately observed the perception of information within audio using various compression formats [2]. Having said this little is known or has been investigated with the primary conjecture proposed in this report.

4. Unreal Engine 4

Unreal Engine is a three-dimensional 3D computer graphics engine built to develop computer games. Unreal engine was created by Epic Games in 1998 [8]. Version 4 or commonly known as Unreal Engine 4 ($UE4$ 4.27), is the fourth iteration of the game development software, and is the version used for the project [9]. UE4 was used as it was able to host development editors for the virtual reality development kit from Oculus. This allows for general 3D graphics to be rendered with a virtual reality headset (Oculus Quest 2) [14]. With the proposed solutions, the

development of an environment used for testing was created using this platform shown in Figure 6.

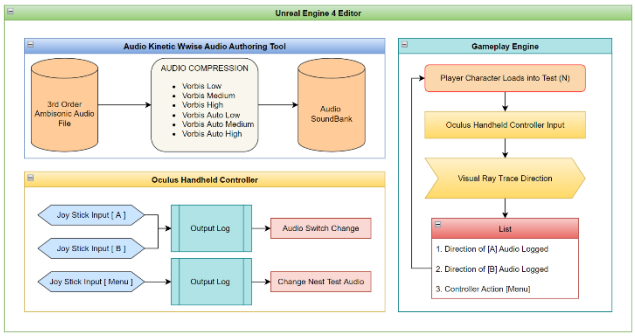

Figure 4. Flow diagram showing the overview of the proposed test, and the individual elements that are present during the test. Wwise integration prerequisite with oculus ambisonic development [12].

5. Audiokinetic Wwise

Audiokinetic Wwise is an audio authoring tool used to bridge saved audio assets and a game engine such as UE4. This software was used as it has the ability to gather positional data from a game engine, and manipulate panning data in real time. The term middleware is commonly used to describe Audiokinetic Wwise, as it theoretically stand between the audio saved in memory and the game engine itself, it does this by using numerous application programming interfaces or commonly known as API's. The current state in the gaming industry shows that Audiokinetic Wwise houses a very large market capitalization boasting partnered relationships with Sony, Microsoft, Android, and Nintendo. [13]. Audiokinetic Wwise version used (2021.1.7.7796) with Wwise > Unreal Engine Integration Version (UEIV) (2021.1.7.7796.2228) [9][13].

6. Test Methodology

6.1 Frequency Specific Reproduction

A general understanding on how the human auditory system and its sensitivity towards specific frequencies, is needed to ensure that the testing audio is suitable, and accurate for the test. Research by Bell Laboratories in 1933 [6]. showed the results of a study that captured audio sensitivity within humans. Shown in Figure 4 is the commonly named Fletcher-Munson Curve (FMC). With this data the conclusion for human sensitivity can be classified as approximately 1 kHz to 5 kHz [6].

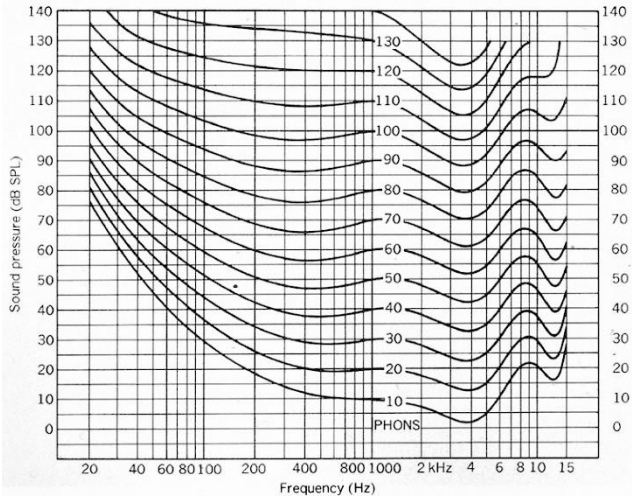

Figure 5. Fletcher-Munson Curve (FMC) showing human sensitivity with respect to frequency and loudness, sound pressure level [6].

As the human auditory system has developed to recognise human voices over general background noise, and the general mean vocal frequency being between 1 kHz to 3 kHz, the test itself will be comprised of human voice, and sonic information within the proposed frequency range [7].

6.2 Test Configuration Virtual Reality

A virtual environment using Unreal Engine 4 and world building development tools were used to generate both a landscape and the testing system used for the listening test. Shown in Figure 6 the generated world environment.

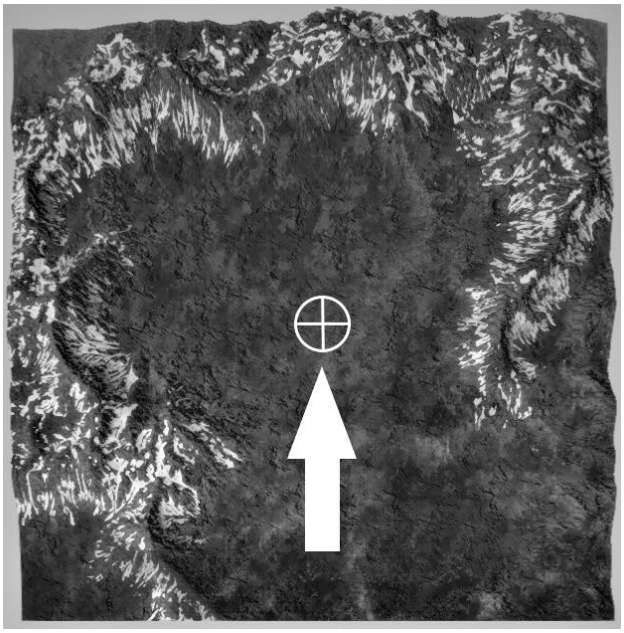

Figure 6. Mountain landscape created in Unreal Engine 4 using world building software, test environment will be set in the geographical centre of the map. Map dimensions (3 Kilometres Length x 3 Kilometres Width) [12].

The use of this map in conjunction with a virtual reality development kit enables the use of the Oculus Quest 2 head mounted display, to view the environment virtually with head tracking [14]. Head tracking was linked to Audiokinetic Wwise via the API call [$WorldPositionPawn$: $BasePlayerActor$]. This allows the head mounted tracking to be correlated in real time to panning coefficients within Wwise [13].

Figure 7. Position of player character VR component used to start the test in the virtual space [12].

Once the environment has been created, the next stage is to define the user inputs for the test, and have the ability to log data. Information about the test sequence within Unreal Engine 4 can be seen in Figure 4.

The Oculus Quest 2 [14] development kit allows the customization of inputs from the device, specifically the hand tracking joystick [14]. This device was used for directional information logging, A to B dynamic change in audio shown in Figure 4, and finally the ability to progress through the test sequence. Shown in Figure 8 is the module coding for the joystick and its fundamental actions within the game engine.

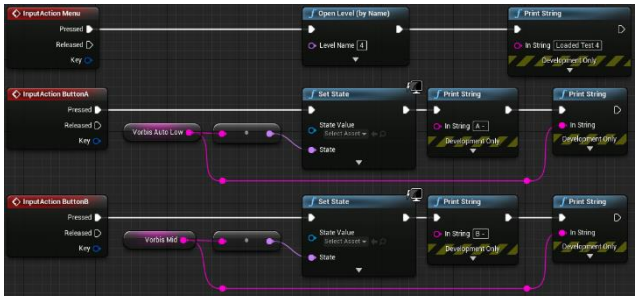

Figure 8. Oculus Quest 2 Right Joystick input parameters, with reference to compression with an AB selection and test progression [12][14].

6.3 Test Configuration Real World

Audio for the test was recorded in an open diffuse environment that was suitable for use within the proposed virtual environment. Source audio was recorded with the ZYLIA-ZM1 3rd order ambisonic microphone. The setup for the microphone in this environment can be seen in figure 9 and 10.

Figure 9. Zylia ZM-1 3rd order ambisonic microphone position within the diffuse environment. This image does not show the microphone connected via USB.

Figure 10. Zylia ZM-1 3rd order ambisonic microphone Connected via USB within the diffuse environment, with oriantation at 0 degrees forward.

6.3.1 Listening Test Setup

The listening test was carried out with the setup shown in Figure 11. Participants were seated in a chair next to the instructor monitoring the test. Participants were then shown the Oculus Quest 2 virtual reality headset that would be worn during the test. Participants were then instructed on the talk of the test, before moving onto a prebuilt virtual environment used as a tutorial session. This session was used to convey information about the test to the participant before the test started, and would tell the participants exactly what was expected of them during the test, as well as health and safety information in the event of an emergency. Once the test had been concluded, the instructor asked the participant what they thought of the test, and if the information on the test was conveyed correctly.

Figure 11. Listening test set up, with Oculus Quest 2 virtual reality headset and joystick.

6.3.2 Listening Test Configuration

The test configuration was set by having the participant subjected to ambisonic audio that was binauralized to stereo over headphones. During the test, the participant would be subjected to audio that had had been set to a specific azimuth, that changed on a pre-determined model seen in Table 2. The participant would then be asked to switch between A or B on the Oculus joystick, corresponding to different compression ratios shown in Table 2, expressed in Figure 1. The selection of A and B audio did not change the time domain of the signal being sent to the participant, only the compression, as the test was set by playing all compressed and non-compressed audio at the same time through the blend function in Audiokinetic Wwise.

		Audio Compression Preset	
Level	Audio Azimuth Degrees From 0	[A]	[B]
Tutorial	0	Original	Vorbis Auto High
Level 1	30	Original	Vorbis Auto Mid
Level 2	-20	Vorbis High	Vorbis Low
Level 3	-20	Vorbis Auto Low	Vorbis Mid
Level 4	-90	Original	Vorbis High
Level 5	40	Original	Vorbis Mid
Level 6	60	Original	Vorbis Low
Level 7	-130	Vorbis Low	Vorbis Auto Low
Level 8	0	Vorbis Mid	Vorbis Auto Mid
Level 9	-80	Vorbis High	Vorbis Auto High

Table 2. Configuring of the listening test with respect to the levels participants were subjected to during the test, and the corresponding Vorbis audio compression of both A and B selection.

6.3.3 Listening Test Participant Process

1. Participant loads into level (N)
2. Participant points joystick in the direction of the audio cue.
3. Participant selects either A or B buttons on the joystick corresponding to audio compression see Figure 2.
4. Participants logged direction with trigger on the oculus joystick showing a distinctive red line corresponding to the direction.
5. After participant concludes with logging direction of both A and B. Participants press next button on joystick to start level ($N + 1$).

7. Analysis of Results

Data from the tests was taken by measuring the angle in degrees from the participant's input. Over 2700 data points were collected during the test in total. The data was first split into groups referencing the number of participants who performed the test. Once this had been done, all points were taken in respect to the inputs of the test, and placed into a data set, where every column represented a participants input in degrees in relation to both the input of A and B audio, as well as the orientation in degrees.

Initially, every column representing the input of the participant was taken and summed to an average weighting. Taking the mean direction data, and comparing this data to the predetermined position shown in Table 2. The resulting data can be seen in Figure 12.

A quick analysis of the graph shown in Figure 12, shows that there does not seem to be any correlation between what the test parameters were set at in relation to direction, and the direction input of the participants. With participants being subjected to no visual cue towards direction initially the assumption that some data points that were outside of 80 degrees should be ignored. This is due to the input error caused by participants using the oculus joystick without accuracy. Using the dataset

collected previously, the input data that was +-80 degrees out of test set parameters was removed. The subsequent data can be seen in Figure 13.

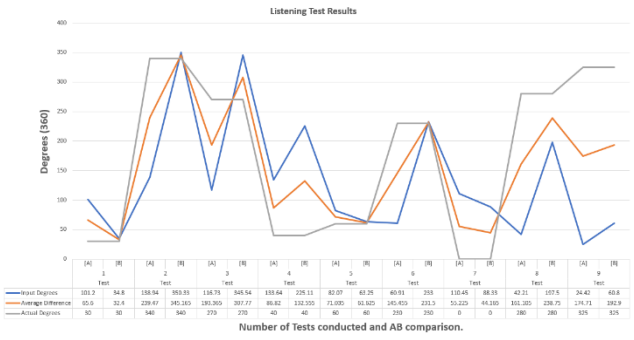

Figure 12. Line Graph showing test data (Blue) compared and averaged (Orange) against the test set parameters (Gray).

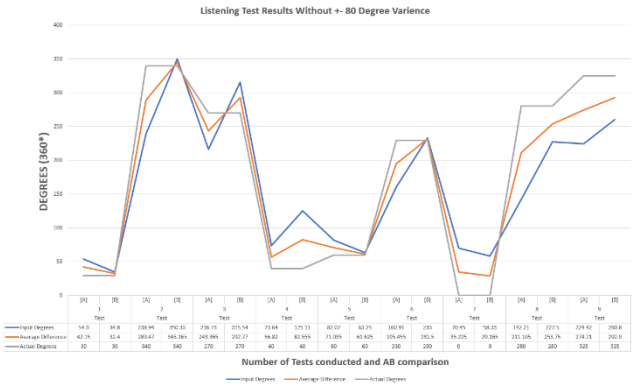

Figure 13. Line Graph showing test data (Blue) compared and averaged (Orange) against the test set parameters (Gray). Without +- 80 Degree Variance.

With the proposed reduction of data within the dataset, the comparison between the test data set parameters, and the participants mean average inputs, show a reduction in variance, which represents a higher accuracy for localization. The problem with this data is that it has been altered and thus can be classified as Bias.

Running a One-Tailed T-Test test to determine difference between data points with a null-hypothesis of (>0.05) as the level of significance. The results of the data can be seen in Table 3, and mathematically shown in 7.1.

With a τ-value of 0.306 and a ρ-value of 0.380426. The results are not significant enough with the set value of $\rho <$ 0.05. "A p-value less than 0.05 (typically $\leq$ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random)." [16]

Treatment 1 (X)	$Diff\,(X - M)$	$Sq.\ Diff\,(X - M)^2$
30	-145.00	21025.00
30	-145.00	21025.00
340	165.00	27225.00
340	165.00	27225.00
270	95.00	9025.00
270	95.00	9025.00
40	-135.00	18225.00
40	-135.00	18225.00
60	-115.00	13225.00
60	-115.00	13225.00
230	55.00	3025.00
230	55.00	3025.00
0	-175.00	30625.00
0	-175.00	30625.00
280	105.00	11025.00
280	105.00	11025.00
325	150.00	22500.00
325	150.00	22500.00
	M: 175.00	SS: 311800.00

Treatment 2 (X)	$Diff\,(X - M)$	$Sq.\ Diff\,(X - M)^2$
54.3	-108.61	11795.53
34.8	-128.11	16411.46
238.94	76.03	5780.98
350.33	187.42	35127.30
216.73	53.82	2896.89
315.54	152.63	23296.76
73.64	-89.27	7968.64
125.11	-37.80	1428.63
82.07	-80.84	6534.66
63.25	-99.66	9931.56
160.91	-2.00	3.99
233	70.09	4913.00
70.45	-92.46	8548.34
58.33	-104.58	10936.40
142.21	-20.70	428.38
227.5	64.59	4172.23
224.42	61.51	3783.82
260.8	97.89	9583.00
	M: 162.91	SS: 163541.55

Table 3. One-Tailed T-Test results for data collected in the listening test. Treatment 1 = pre-set azimuth parameters. Treatment 2 = Mean user azimuth input with relation to Treatment 1. [15]

7.1 Calculation of τ-value

Difference Score Calculations:

Table 1 – Test Set Parameters

N_1: 18
df_1: $N - 1 = 18 - 1 = 17$
M_1: 175
SS_1: 311800

$$s^2{}_1 = \frac{SS_1}{(N-1)} = \frac{311800}{(18-1)} = 18341.18$$

Table 2 – Participant Input Data

N_1: 18
df_1: $N - 1 = 18 - 1 = 17$
M_1: 162.91
SS_1: 163541.55

$$s^2{}_1 = \frac{SS_1}{(N-1)} = \frac{163541.55}{(18-1)} = 9620.09$$

T-Value Calculation:

$$s^2{}_p = \left(\left(\frac{df_1}{(df_1 + df_2)}\right)s^2{}_1\right) + \left(\left(\frac{df_2}{(df_2 + df_2)}\right)s^2{}_2\right)$$

$$= \left(\left(\frac{17}{34}\right)18341.18\right) + \left(\left(\frac{17}{34}\right)9620.09\right) = 13980.63$$

$$S^2{}_{M_1} = \frac{S^2{}_P}{N_1} = \frac{13980.63}{18} = 776.7$$

$$S^2{}_{M_2} = \frac{S^2{}_P}{N_2} = \frac{13980.63}{18} = 776.7$$

$$\tau = \frac{\left(\frac{M_1}{M_2}\right)}{\sqrt{\left(S^2{}_{M_1} + S^2{}_{M_2}\right)}} = \frac{12.09}{\sqrt{1553.4}} = 0.31$$

8. Conclusion

Although the test results shown in Figures 12. 13. show a promising correlation that ambisonic audio that has been compressed with the lossy compression Vorbis, does not show variance in localization with respect to compression ratio, the addition of the T-Test statistically shows that there is no significant data to prove this hypothesis, thus meaning that there cannot be any distinguishable conclusions based on the data shown in this report. This is due to the lack of data points for analysis, and the apparent error inherited by the equipment used in the test.

9. References

[1] *Teddy Surya Gunawan, Siti Aisyah Abdul Rashid, Mira Kartiwi. (2017). Investigation of Various Algorithms on Multichannel Audio Compression. IEEE International Conference on Smart Instrumentation, Measurements and Applications (ICSIMA) . 4 (3), 3.*

[2] *Erik Montnémery Johannes Sandvall. (2004). Psycho acoustic audio coding. In: Erik Montnémery Johannes Sandvall Psycho acoustic audio coding. Lunds Universitet: Lunds Universitet. 11.*

[3] *Henrique S. Malvar (1992). Signal Processing with Lapped Transforms. United States of America: Artech House, Inc 314.*

[4] *Audiokinetic. (2022). Integrations. Available: https://www.audiokinetic.com/en/. Last accessed 10/05/2022.*

[5] *Wiggins, B. (2017) 'Analysis of binaural cue matching using ambisonics to binaural decoding techniques' Presented at 4th International Conference on Spatial Audio, Graz, Austria, 7th-10th September.*

[6] *Harvey Fletcher, Wilden A. Munson. (1933). Loudness, Its Definition, Measurement and Calculation. : The Journal of the Acoustical Society of America. 148 (4), 91.*

[7] *Amy D. Bagley, Carolyn S. Abramowitz, David S. Kosson. (2009). Vocal Affect Recognition and Psychopathy: Converging Findings Across Traditional and Cluster Analytic Approaches to Assessing the Construct. Journal of Abnormal Psychology. 118 (2), 388.*

[8] *Mike Thomsen. (2012). History of the Unreal Engine. Available: https://www.ign.com/articles/2010/02/23/history-of-the-unreal-engine. Last accessed 15/05/2022.*

[9] *Epic Games. (2022). Unreal Engine 4. Available: https://docs.unrealengine.com/4.27/en-US/. Last accessed 15/05/2022.*

[10] *Xiph.Org Foundation. (1994). Vorbis I specification. Available: https://xiph.org/vorbis/doc/. Last accessed 15/05/2022.*

[11] *Microsoft. (2022). Durango Audio Engine Developer Doc. ID@XBOX . 1 (-), -.*

[12] *Self. (2022) EFFECTS OF LOSSY COMPRESSION CODECS ON THE PERCEPTION OF IMMERSIVE AUDIO IN UNREAL ENGINE 4.*

[13] *Audiokinetic. (2022). Release Notes 2021.1.8 . Available: https://www.audiokinetic.com/library/edge/?source=SDK&id=releasenotes.html. Last accessed 15/05/2022.*

[14] *Meta, Oculus. (2022). Unreal Engine. Available: https://developer.oculus.com/documentation/unreal/. Last accessed 15/05/2022.*

[15] *socscistatistics. (2022). T-TEST. Available: https://www.socscistatistics.com/tests/studentttest/default2.aspx. Last accessed 18/05/2022.*

[16] *Dr. Saul McLeod. (2019). What a p-value tells you about statistical significance. Available: https://www.simplypsychology.org/p-value.html#:~:text=A%20p%2Dvalue%20less%20than,and%20the%20results%20are%20random).. Last accessed 18/05/2022.*