OPEN ACCESS Freely available online

B. Lee, T. Rudzki, J. Skoglund, and G. Kearney, "Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality"
J. Audio Eng. Soc., vol. 71, no. 4, pp. 145–154, (2023 April).
DOI: https://doi.org/10.17743/jaes.2022.0068

Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality

BEN LEE,[1] *AES Associate Member*, **TOMASZ RUDZKI,**[1] *AES Student Member*,
(ben.lee@york.ac.uk) (tomasz.rudzki@york.ac.uk)
JAN SKOGLUND,[2] *AES Member* **AND GAVIN KEARNEY,**[1,*] *AES Member*
(jks@google.com) (gavin.kearney@york.ac.uk)

[1]*AudioLab, School of Physics, Engineering and Technology, University of York, York, United Kingdom*
[2]*Google LLC., San Francisco, CA*

This paper discusses the evaluation of Opus-compressed Ambisonic audio content through listening tests conducted in a virtual reality environment. The aim of this study was to investigate the effect that Opus compression has on the Basic Audio Quality (BAQ) of Ambisonic audio in different virtual reality contexts—gaming, music, soundscapes, and teleconferencing. The methods used to produce the test content, how the tests were conducted, the results obtained and their significance are discussed. Key findings were that in all cases, Ambisonic scenes compressed with Opus at 64 kbps/ch using Channel Mapping Family 3 garnered a median BAQ rating not significantly different than uncompressed audio. Channel Mapping Family 3 demonstrated the least variation in BAQ across evaluated contexts, although there were still some significant differences found between contexts at certain bitrates and Ambisonic orders.

0 INTRODUCTION

Over the past decade, the rise of streaming services like Netflix and Spotify has led to a paradigm shift in the methods used by people to access and digest digital media [1]. Media is now widely accessible globally and available almost instantly. Moreover, not only has there been a shift in the way people access content, but also a shift in the actual production and reproduction of the content itself. Innovations in spatial audio in particular have garnered a certain momentum in recent years.

With the increasing popularity of more complex and data-heavy media formats, coupled with the rising demand for instantaneous content, work needs to be done to ensure that content streaming is as efficient as possible. Couple this increasing popularity with the growing sophistication and interest in virtual reality (VR) systems, and there becomes not only a reason but an imperative to ensure that the components that facilitate the quality of VR systems are evaluated and improved. Spatial audio plays a significant role in the standard of VR, and without it, the environments would not be nearly as immersive.

*To whom correspondence should be addressed, e-mail: gavin.kearney@york.ac.uk

However, there are many different contexts of media consumption in VR, and content can vary from simple 360° non-interactive soundscapes to fully interactive occupational training simulations. Context may therefore dictate more sophisticated spatial audio processing, e.g., higher bitrates of compression or certain compression algorithms.

Opus is one such algorithm that caters to a wide variety of audio applications, from voice-over IP to streaming live music performances. The use of Opus for immersive audio has been studied previously using standard, 2D screen-based data collection methods, without any accompanying visual stimulus. Studies by Narbutt et al. [2, 3] focused on two perceptual attributes: Listening Quality and Localization Accuracy for Opus-compressed first-order and third-order Ambisonic stimuli in order to facilitate the development of the AMBIQUAL objective audio quality metric. These tests were carried out using static binaural rendering employing generic head-related transfer functions (HRTFs). The tested version of the codec was Opus 1.2 employing Channel Mapping Family 2. Another study by [4] focused on the evaluation of the codec using loudspeaker-based rendering. The tested perceptual attribute was Timbral Distortion. The Opus codec used in that study employed Channel Mapping Family 3. A brief characterization of Opus channel mapping families is provided in SEC. 1.5.

1 METHODOLOGY

This section presents the methods used to design and implement the multi-stimulus listening test, including content creation. The study builds on the ITU-R BS.1534-3 [Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)] [5] guidelines but does not strictly follow these recommendations because they were never designed for head-tracked VR presentations. For instance, listening to a hidden reference while looking in a different direction to the actual reference could cause listeners to rate this condition lower even with no degradation invoked by Opus compression.

1.1 Content Rationale

Four different contexts were evaluated: gaming, music, soundscapes, and teleconferencing. Firstly, gaming is arguably the most common scenario in which VR is experienced, demonstrated by the sheer amount of VR headsets marketed as accessories to PlayStation [6], or even as an independent, all-in-one, gaming console [7]. Ambisonics has been proven to be useful for game audio, e.g., in racing simulators [8]. Opus can be used for compressing game assets and for streaming networked virtual gaming/VR experiences, such as interactive Metaverse concerts or viewing of e-Sports, not to mention the widespread streaming of recorded game content on YouTube.

Music is another context in which spatial audio has entered the market. For example, Apple recently released support on their music streaming platform for the Dolby Atmos spatial audio format [9]. Music listening is therefore likely to become a more prevalent context within VR.

Soundscapes are another important context to evaluate. Not only have they been used in previous studies and evaluations of spatial audio in conjunction with VR [10], but soundscape-based virtual environments have been used in health and well-being applications, for example, with children who have Autism Spectrum Disorder (ASD) for behavioral response measurement [11], therapeutic treatment [12], disaster awareness [13], and as educational tools [14].

Teleconferencing has become an integral part of everyday life for many people in the years since the COVID-19 pandemic began. "Zoom, for instance, had 10 million daily meeting participants in December 2019, but by April 2020, that number had risen to over 300 million" [15]. Accompanying this shift in the way colleagues interact are many changes and obstacles that impede the efficiency and naturalness of conversation. For instance, latency can lead participants in a teleconference to talk over each other, and a lack of any sort of spatial dimension or feeling as though other participants are not sharing the same physical space can cause other problems, such as distraction or disinterest.

Karl et al. discuss the Media Naturalness Theory, in which five key characteristics of media naturalness are presented. VR teleconferencing may help improve all of these areas to some degree. The use of spatial audio in VR for teleconferencing could help to improve co-location by immersing participants within the same perceived environment, VR environments could help with facial expression and body language by giving participants full-body avatars, spatial audio could also help with speech intelligibility by making it easier to discern who is talking, and, finally, the issue of synchronicity can be improved by studies, such as this one, in order to make the streaming of Ambisonic audio as efficient as possible.

1.2 Production of Visual Content

Despite the fact that the test scenes represented the different contexts of gaming, music, soundscapes, and teleconferencing, to facilitate a fair comparison between the contexts and test conditions, the material had to be pre-rendered. This meant that apart from three-degrees-of-freedom, head-tracked rotation of the scenes, no other interaction was facilitated. In other words, participants could not actually play the game in game scenes, play instruments or sing in music scenes, or speak or interact with the avatars in teleconference scenes. Therefore, all of the scenes were 360° videos of the specific contexts they were representing in which the MUSHRA-style test interface could be overlaid.

The foundations of said scenes were created using the Unity platform.[1] Once a Unity project was made for each scene, with corresponding sound-emitting objects, scripts were added to the objects which produced JSON files. The JSON files captured information about the sound-emitting object such as positions relative to the listener at time intervals. This information was needed in order to spatialize the audio for each object accurately after the scene was created. The scene could then be played in Unity and recorded using Unity's proprietary 360° video recorder whilst also producing JSON files for each of the moving objects.

1.3 Description of Scenes

Two different scenes were made for each of the four contexts. All scenes had a duration of 12 s and are described as follows:

- *GameCar*: First-person perspective of a driving game in which the participant is in the driving seat. There are different colored cars that race while colliding with each other and the walls.
- *GameFPS*: First-person shooter style game in which the participant is in a room with an alien and UFO as enemies; a gun can be seen shooting the enemies until they are destroyed.
- *MusicBlues*: Scene in which the participant is in a room surrounded by various sounding instruments, e.g., piano, brass, and percussion.
- *MusicMallets*: Scene in which the participant is in a room surrounded by various sounding instruments, e.g., keyboard, vibraphone, strings, and percussion.
- *SoundscapeFarm*: Scene in which the participant is in a farmer's field, surrounded by various low poly objects: environmental objects, stationary sound emitting farm animals, and a moving tractor.

[1] Unity: https://unity.com/.

- *SoundscapeOasis*: Scene in which the participant is in a desert oasis. The participant is stood next to water and camels; there are also some nearby palm trees and a propeller plane passing in the distance.
- *TeleconferenceOne*: Scene in which the participant is at a table, with four animated mannequin avatars.
- *TeleconferenceTwo*: Scene in which the participant at a table, with three animated mannequin avatars.

The test soundtracks were intentionally not made overdense with sound effects to reduce the cognitive load of the user and allow them to better focus on the timbre and spatial positioning of sounds. To reduce test duration and potential listener fatigue, scenes were divided between two test sessions. Each scene was spatialized using first-order, third-order, and fifth-order Ambisonics, and each Ambisonic order was assessed separately. Therefore, Test Sessions 1 and 2 comprised 12 trials each.

1.4 Production of Audio Content

Because first-order, third-order, and fifth-order Ambisonic audio is evaluated in this study, virtually produced Ambisonic tracks were made via spatialization of mono tracks. Without up-mixing lower-order Ambisonic content, this was the only way to produce original fifth-order Ambisonic content for the VR scenes. Actually capturing fifth-order Ambisonic content from the real world was not possible because fifth-order higher-order Ambisonics microphones were not commercially available.

The 360° videos of the scenes were rendered using Unity's video recorder; no audio was rendered in Unity because of its limitations in the binaural rendering of higher-order Ambisonic signals. This is why the JSON files, containing positional information for each object, were necessary to be exported along the video. Once a mono audio track was created for each of the objects in a scene, this positional data—from the respective JSON file—could then be used to spatialize the track using Ambisonics.

The mono tracks were produced as WAV files at a 48-kHz sample rate and bit-depth of 16 bits. Some of them were produced entirely with VSTs in Logic Pro X, and others incorporated royalty-free sounds. All of the test stimuli that were produced for this study, Ambisonic audio files, 360° videos, and a track list of sound effects for all of the outsourced pieces of audio used when arranging the mono audio tracks, have been made publicly available for download [16].

In order to read the mono WAV files and corresponding JSON files for each object, a MaxMSP [17] patch was used. MaxMSP objects can be used to run plug-ins, and this was necessary in order to spatialize the mono tracks through the use of IEM's Ambisonic plug-in suite [18]—"Room Encoder" allowed for the simulation of early reflections, and the "Stereo Encoder" for Ambisonic encoding.

First-order, third-order, and fifth-order Ambisonic scene files were created for each of the eight scenes mentioned in SEC. 1.3. These Ambisonic WAV files would then go on to be the reference audio from which the Opus-compressed files would be created.

Before compressing the Ambisonic files created in Max MSP, each file was adjusted in MATLAB so that the perceived volume of each scene was the same in Loudness Units Full Scale (LUFS). The AmbiX v0.2.10 – Ambisonic plug-in suite [19] was used to binauralize the Ambisonic files, along with head-related impulse responses and binaural decoding configuration files taken from the SADIE II database [20, 21]; these matched the target rendering chain, assuming no head rotation, during the experiment, which is described in SEC. 1.8.

Although MATLAB can analyze the LUFS of multichannel files, there is no standard way to determine the weighting of Ambisonic components and their contribution to the loudness. It was therefore deemed more appropriate to analyze the LUFS of each scene rendered binaurally using the same rendering workflow as in the listening test. An arbitrary value of –31 LUFS was used as the target level for all of the binaural renderings. A gain compensation was calculated for each and then applied to the Ambisonic files.

1.5 Opus Compression and Anchor Creation

The Opus parameters tested were bitrate and channel mapping family. Channel Mapping Family 2 codes each Ambisonic component as an independent Opus stream, i.e., each of the channels are encoded separately. This direct uncoupled method does not take advantage of useful codec features such as coupled stereo mode. Channel Mapping Family 3 does utilize these features through projection-based compression in which channels are effectively coupled together in pairs when coded, and then a demixing matrix is used to separate the channels upon decoding. A more detailed description of each of these channel mapping families can be found at [22]. The publicly available version of Opus contains the first-order and third-order Ambisonics matrices for Channel Mapping Family 3; a patch was added to the official Opus 1.3.1 code to support fifth-order Ambisonics, which can be provided upon request.

Each Ambisonic reference file was compressed at 16, 32, and 64 kilobits per second per channel (kbps/ch) and for both channel mapping families, 2 and 3. This meant that the total amount of compressed files, for each scene, at each Ambisonic order, was six. These six different compression conditions composed each trial, along with a hidden reference and mid-range and low anchors. The low anchor was a low-pass–filtered version of the reference Ambisonic audio with a cut-off frequency of 3.5 kHz; the mid-range anchor had a cut-off frequency of 7 kHz. Therefore, each trial comprised of nine different conditions.

1.6 Listening Test Environment

For this experiment, the Spatial Audio Listening Test Environment (SALTE) software was employed [23]. The software consists of a dedicated standalone app containing test control and binaural audio rendering modules. A sepa-

rate app, SALTE for VR,[2] was used to render visual content on a stand-alone VR headset (Oculus Quest 2).

1.7 Test Setup

To set up a MUSHRA test in SALTE, a configuration JSON file is used to specify the reference and condition audio files for each trial and to set other parameters of the test. This includes any additional gain applied to the audio files, whether headphones are used instead of loudspeakers, which HRTFs will be used for the subsequent binauralization, and the location where these HRTFs can be found on the disk. The 360° video file name for each trial is also set in the configuration file. Once the correct configuration file was selected, the participant had to take note of their randomized subject ID which was used as an anonymous marker. The test results were exported into a single CSV file.

1.8 Binaural Rendering

The audio rendering module of SALTE was configured to process the Ambisonic audio and output binaural renderings of each scene while simultaneously processing head-tracking data from the VR headset so that the Ambisonic scene could be rotated in real time. Three different virtual loudspeaker configurations, octahedron, 26-point Lebedev grid, and 50-point Lebedev grid, were employed for the rendering of first-order, third-order, and fifth-order Ambisonics, respectively [24]. Dual-band decoding was implemented by pre-filtering the Ambisonic input with a set of shelf filters[3] and applying Max-Re correction weightings to the high-passed signals before feeding the decoder. Ambisonic decoder configuration files were obtained from the SADIE II database [20, 21]. To create binaural signals, loudspeaker feeds were convolved in real time with diffuse-field equalized KU100 HRTF sets obtained from the SADIE II database.

1.9 VR Interface

Oculus Quest runs the Android operating system so the final step was to produce an Android application package (APK), which could be installed onto the headset to display 360° videos and connect to the SALTE desktop program wirelessly; there is no software that currently allows the Quest to render higher-order Ambisonic audio and 360° video simultaneously, so SALTE ran on a desktop computer as the sole audio renderer. Therefore, the VR app needed to send Open Sound Control data containing head-tracking information from the headset to the desktop SALTE audio renderer so that the Ambisonic scene could be counter-rotated. Other information, such as the ratings for conditions and the actual test interface display, needed to be passed between the desktop and headset so that they could move and respond in coordination.

The 360° video also needed to be in sync with any audio being output by SALTE—this was achieved by measuring the time taken for video playback to start and by delaying the audio playback by the same amount of time so that both played simultaneously. This was one of the reasons that the playback interval could not be changed by the participant in real time because it could have caused video and audio to become out of sync.

The SALTE for VR program was developed using Unity. All of the 360° video content had to be uploaded to the Unity project in a Streaming Assets folder so that it would be installed onto the headset and therefore directly accessible by the headset when rendering the visuals. The APK file could then be built and uploaded to any Oculus Quest using SideQuest [25] using a USB-C cable.

Once these settings were configured for the SALTE desktop program, the participant had to put their Oculus Quest on, run the APK file, and follow the on-screen instructions. The IP address of the participant's Quest is shown, and this is then input into the SALTE desktop program to link the SALTE desktop and SALTE for VR programs together. Once the headset and computer were linked, the participant could begin the test.

Upon clicking "Begin," the participant was presented with the test-rating interface overlaid onto the 360° video for the first trial. The participant was then able to click on each of the conditions to listen to them. Pressing on the corresponding playback buttons would play, pause, or stop the audio. The sliders, above each of the conditions, could be dragged up and down to rate the corresponding condition between 0–100 based on how similar it was to the reference audio. The participant would move between all of the trials using the "Next" or "Previous" buttons until they were happy with the ratings they had given to each of the conditions in every trial. A "Finish" test button would then appear on the final trial, which, upon clicking, ended the test. Fig. 1 shows the SALTE for VR test interface with all of the described features, such as sliders and playback buttons.

1.10 Test Procedure

The listening tests were originally planned for distribution to participants so that they could be completed in the comfort of their own homes. However, shortly after the initial test distribution, COVID-19 restrictions had relaxed enough for the remainder of the tests to be set up and completed at the University of York's AudioLab. This meant that the remainder of the participants no longer had to download, install and set up the SALTE and SALTE for VR programs on their home devices; this streamlined the test process and allowed for a more efficient collection of a larger quantity of data sets from multiple participants.

1.11 Data Analysis

To reduce the number of independent variables so as to make the analysis less convoluted, the data collected using stimuli encoded for Ambisonic orders was assessed sepa-

[2]https://github.com/trsonic/SALTE4Quest-XRIT.

[3]https://github.com/resonance-audio/resonance-audio/tree/master/matlab/ambisonics/shelf_filters.

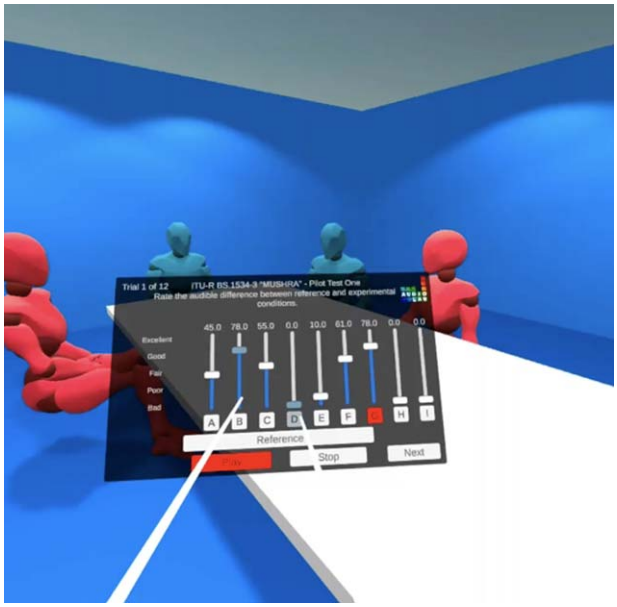

Fig. 1. A screenshot of the SALTE for VR MUSHRA test interface overlaid on top of the TeleconferenceOne scene.

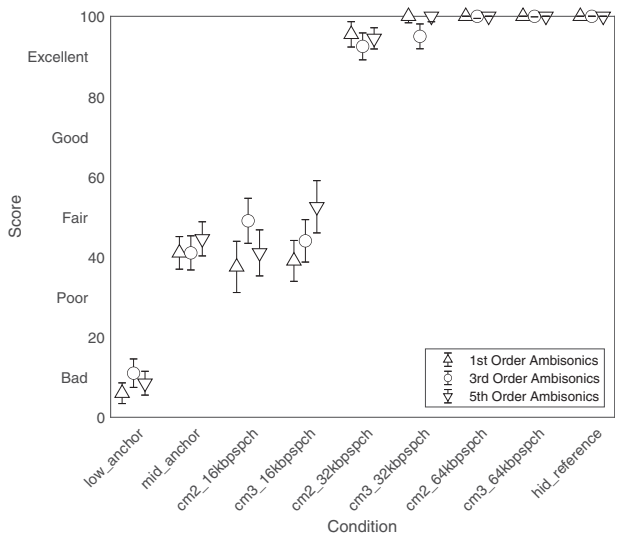

Fig. 2. Median BAQ ratings for all conditions aggregated over all contexts. Whiskers denote 95% confidence intervals.

rately. This study does not address differences in perceived audio quality between different Ambisonic orders.

The first null hypothesis to be investigated is that there there is no significant difference between the Basic Audio Quality (BAQ) of compressed and uncompressed stimuli. The second null hypothesis is that, for each bitrate of compression, the channel mapping family has no effect on the BAQ rating. The independent variable in this example is the channel mapping family, within a certain bitrate of compression, and the dependent variable is the BAQ rating. The third null hypothesis is that for each of the codec parameters, the BAQ rating is independent of scene context.

Expert listeners were employed as subjects. To ensure that the listeners were suitable candidates, people with extensive listening test experience, i.e., professionals and Ph.D. students who work in the field of music and/or audio were chosen to complete the tests. The following criteria were applied in the post-screening of participants: Responses of the assessor collected within a single listening test session were excluded if they rated the hidden reference condition for more than 20% of the test items lower than a score of 90 or if they rated the mid-range anchor for more than 33% of the test items higher than a score of 95.

The responses collected in both listening test sessions were analyzed separately with the above exclusion principles being applied. Therefore, some participants might have been excluded, e.g., from the first session but not the second one. These criteria resulted in the exclusion of seven assessors due to missed hidden reference and further four assessors due to mid-range anchor rated too high.

1.12 Test for Normality

The Shapiro-Wilk test was used in order to determine whether the data was normally distributed. All data was divided into 24 Order-Scene groups. Each group contained ratings for each of the conditions; as mentioned in SEC. 1.5, there were a total of nine conditions in each trial, meaning there were 216 different distributions tested using the Shapiro-Wilk test. Out of the 216 distributions that were analyzed, 150 were not normally distributed. Therefore, it was determined that non-parametric statistical analysis will be conducted using the Kruskal-Wallis test and subsequent comparison of mean ranks.

2 RESULTS

The conditions in the figures are labeled as "cm2_16kbpspch," "cm2_32kbpspch," "cm2_64kbpspch," "cm3_16kbpspch," "cm3_32kbpspch," "cm3_64kbpspch," "hid_reference," "low_anchor," and "mid_anchor," corresponding respectively to the following conditions, also described in SEC. 1.5: 16 kbps/ch Channel Mapping Family 2, 32 kbps/ch Channel Mapping Family 2, 64 kbps/ch Channel Mapping Family 2, 16 kbps/ch Channel Mapping Family 3, 32 kbps/ch Channel Mapping Family 3, 64 kbps/ch Channel Mapping Family 3, Hidden Reference, Low anchor, and Mid-range anchor.

2.1 General Comparison

Fig. 2 shows median BAQ ratings aggregated over all eight scenes, differentiated by Opus codec parameters and Ambisonic orders. For each test scene, the conditions were evaluated against a reference track rendered at the same Ambisonic order; therefore, differences in perceived BAQ between different Ambisonic orders are not revealed. The non-parametric 95% confidence intervals have been computed based on the standard formula used to calculate the size of a notch in a boxplot [26].

It can be seen that for every Ambisonic order, the low anchor garnered the lowest BAQ rating of all conditions. The mid-range anchor, 16 kbps/ch Channel Mapping 2, and 16 kbps/ch Channel Mapping 3 conditions all received the

next-lowest BAQ rating, and there were no significant differences between these conditions at respective Ambisonic orders. The 32 kbps/ch Channel Mapping 2 conditions received lower median BAQ values for all Ambisonic orders than 32 kbps/ch Channel Mapping 3 conditions. However, the only significant difference between these two conditions can be seen for the fifth-order Ambisonics, because the whiskers do not overlap. This helps to disprove the second null hypothesis, that the use of Channel Mapping Families 2 and 3 results in the same perceived quality of encoded audio. There were no significant differences at any Ambisonic order between 64 kbps/ch Channel Mapping 2, 64 kbps/ch Channel Mapping 3, and the hidden reference conditions—partially supporting the first null hypothesis. In general, Channel Mapping 3 garnered higher BAQ ratings in most instances, except for third-order Ambisonic scenes compressed at 16 kbps/ch.

2.2 Perceived Audio Quality Impairment

This section addresses the problem of finding codec parameters that do not cause perceived degradation of BAQ, through a comparison of all experimental condition scores against uncompressed stimuli scores. Table 1 shows p values obtained from pairwise comparisons between hidden reference and each of the remaining conditions using the Wilcoxon rank sum test. Each of the bitrates is analyzed to test the null hypotheses separately.

2.2.1 64 kbps/ch Bitrate

It can be seen from Table 1 that there are no significant differences present between the hidden reference and any of the stimuli compressed at 64 kbps/ch using Channel Mapping Family 3. For some of the stimuli, the ratings are even significantly the same as for the hidden reference, suggesting perceptual transparency of Opus compression at these settings. This finding, therefore, supports the first null hypothesis that there is no significant difference between the median BAQ rating of the uncompressed reference and Opus-compressed stimuli at the 64-kbps/ch bitrate using Channel Mapping Family 3.

The ratings of stimuli encoded at 64 kbps/ch using Channel Mapping Family 2 follow the same trend in many cases because there is only one trial of the 24 in which the median BAQ rating was significantly different than the hidden reference rating. This finding therefore also supports the first null hypothesis in all trials but the Music Blues scene encoded using third-order Ambisonics.

2.2.2 32-kbps/ch Bitrate

The 32-kbps/ch compressed material for both channel mapping families had instances in which it was significantly different than the hidden reference, disproving the first null hypothesis. This bitrate showed greater variation between channel mapping families than the other bitrates. To elaborate, Channel Mapping Family 2 was significantly different than the hidden reference more times than its counterpart, Channel Mapping Family 3. At this bitrate Channel Mapping Family 2 was significantly different than the hidden reference in 19 of the 24 trials, whereas Channel Mapping Family 3 was significantly different than the hidden reference in 10 of the 24 trials. This finding suggests that there is a difference between channel mapping families, thereby rejecting the second null hypothesis.

2.2.3 16-kbps/ch Bitrate

The 16-kbps/ch compressed material was significantly different than the hidden reference in all cases but one, in which Channel Mapping Family 2 showed neither significance for the Game Car first-order Ambisonic scene. The first null hypothesis is therefore disproven at this bitrate in most cases because there was a significant difference between the compressed stimuli and hidden reference. These results also support the second null hypothesis at this bitrate because both channel mapping families were rated significantly different than the hidden reference in all trials but one, which was the only difference between the channel mapping families at this bitrate. Whether the second null hypothesis is actually proven at this bitrate is discussed in SEC. 2.3.

2.2.4 Other Findings

An interesting finding that Table 1 presents clearly is that the mid-range anchor was never significantly different and even rated significantly the same as the hidden reference twice, across the trials utilizing the Game Car scene. This makes sense upon reflection because the uncompressed reference audio consisted of low-pass–filtered audio for engine and road sounds typically found in first-person car games. The mid-range anchor applies a 7.5-kHz low-pass filter to the uncompressed audio which would have had little to no effect if the audio was already low pass filtered below this frequency during the stimuli production phase.

At first glance, it is difficult to tell whether certain contexts were affected differently by Opus compression but it can be seen from Table 1 that certain contexts had more conditions with median BAQ ratings significantly different than the hidden reference than other contexts. This contradicts the third null hypothesis that for each of the codec parameters, the BAQ rating is independent of the scene context. The context where most conditions were rated differently from the hidden reference was music, which could suggest that this content is more difficult to encode than others. Further analysis is required to fully determine whether or not the third null hypothesis has been proven or disproven by these results.

2.3 Channel Mapping Family Effect on BAQ

This section addresses whether there are perceived differences in BAQ ratings between Channel Mapping Families 2 and 3 at each bitrate. Table 2 shows p values obtained from pairwise comparisons between the two channel mapping families for all contexts combined at separate bitrates and Ambisonic orders using the Wilcoxon rank sum test.

For 64 kbps/ch, the second null hypothesis holds true because no significant difference could be found between the two channel mapping families at this bitrate at any Am-

Table 1. The p values obtained from pairwise comparisons between uncompressed audio (hidden reference) and each of the remaining conditions using Wilcoxon rank sum test. Assuming a p-value threshold of 0.05, the significantly different rating score distributions are marked by p-values in bold.

	low_anchor	mid_anchor	cm2 _16kbpspch	cm3 _16kbpspch	cm2 _32kbpspch	cm3 _32kbpspch	cm2 _64kbpspch	cm3 _64kbpspch
GameCar_1OA	**0.000053**	0.980273	0.079739	**0.00023**	0.718707	0.663576	0.754917	0.791644
GameCar_3OA	**0.000036**	1	**0.000227**	**0.000031**	0.399109	0.862624	0.798117	0.791674
GameCar_5OA	**0.000878**	0.138096	**0.001282**	**0.000002**	0.260397	**0.017076**	0.455617	0.695352
GameFPS_1OA	**0**	**0**	**0.000008**	**0.000007**	0.11014	0.769304	0.777101	0.961729
GameFPS_3OA	**0**	**0**	**0.000004**	**0.000001**	**0.001556**	0.076657	0.279069	0.536784
GameFPS_5OA	**0**	**0**	**0.000001**	**0**	**0.007141**	**0.042497**	0.346427	0.097076
MusicBlues_1OA	**0.000001**	**0.000001**	**0.000028**	**0.000006**	**0.014397**	**0.043275**	0.916158	0.632725
MusicBlues_3OA	**0.000001**	**0.000001**	**0.000028**	**0.000028**	**0.008258**	**0.00152**	**0.018066**	0.079725
MusicBlues_5OA	**0.000001**	**0.000001**	**0.000021**	**0.000004**	**0.014247**	**0.000259**	0.325821	0.179751
MusicMallets_1OA	**0**	**0**	**0**	**0**	**0.025387**	**0.003929**	0.422191	0.910876
MusicMallets_3OA	**0**	**0**	**0**	**0**	**0.000275**	0.076628	0.572773	1
MusicMallets_5OA	**0**	**0**	**0**	**0.000001**	**0.003531**	0.257501	0.412137	0.45305
SoundscapeFarm_1OA	**0.000002**	**0.000017**	**0.000096**	**0.000019**	**0.017676**	0.663629	0.837836	0.978746
SoundscapeFarm_3OA	**0.000001**	**0.000068**	**0.000007**	**0.000129**	**0.000876**	0.092739	0.151459	0.48852
SoundscapeFarm_5OA	**0.000002**	**0.000215**	**0.000004**	**0.000018**	**0.048918**	0.215485	0.942572	0.823946
SoundscapeOasis_1OA	**0**	**0**	**0.000001**	**0**	**0.000063**	0.295024	0.860367	1
SoundscapeOasis_3OA	**0**	**0**	**0.000003**	**0**	**0.000435**	**0.023186**	0.740057	0.334596
SoundscapeOasis_5OA	**0**	**0**	**0**	**0**	**0.000005**	**0.000001**	0.178139	0.916615
TeleconferenceOne_1OA	**0.000001**	**0.000001**	**0.000001**	**0.000001**	**0.019058**	**0.006702**	0.361567	0.361415
TeleconferenceOne_3OA	**0.000002**	**0.000002**	**0.000002**	**0.000011**	**0.034945**	0.10601	0.910914	0.918453
TeleconferenceOne_5OA	**0.000001**	**0.000001**	**0.000001**	**0.000006**	**0.003675**	**0.013402**	0.389423	0.609312
TeleconferenceTwo_1OA	**0**	**0**	**0**	**0**	**0.019373**	0.083438	0.166298	0.153641
TeleconferenceTwo_3OA	**0**	**0**	**0**	**0.000001**	**0.046467**	0.402262	0.211677	0.322942
TeleconferenceTwo_5OA	**0**	**0**	**0**	**0**	0.414746	0.405687	0.701748	0.789881

Table 2. The p values obtained from pairwise comparisons between Channel Mapping Families 2 and 3 using Wilcoxon rank sum test. Assuming a p-value threshold of 0.05, the significantly different rating score distributions between the both channel mapping families are marked by p-values in bold.

	16 kbps/ch	32 kbps/ch	64 kbps/ch
1OA	0.474799	**0.007192**	0.951408
3OA	**0.034928**	**0.001441**	0.112376
5OA	0.373840	0.555900	0.516494

bisonic order—first-order Ambisonic content at this bitrate even showed that the channel mapping families were significantly the same. For 32 kbps/ch, the second null hypothesis is rejected for first-order and third-order Ambisonics because there was a significant difference between the channel mapping families—no significance, either way, was determined at this bitrate for fifth-order Ambisonics. Finally, for 16 kbps/ch, the second null hypothesis is rejected for third-order Ambisonics because there was a significant difference between the channel mapping families—no significance, either way, was determined at this bitrate for first-order and fifth-order Ambisonics.

2.4 Stimulus Context Effect on BAQ

This section addresses the problem of determining whether there is any difference in BAQ ratings between contexts at each bitrate. Each condition from each of the trials was combined except for the separate Ambisonic orders. Table 3 shows p values obtained using the Kruskal-Wallis test for all contexts at separate conditions and Ambisonic orders. Each of the bitrates was analyzed and used to test the third null hypothesis separately.

For 64 kbps/ch Channel Mapping Family 2, the third null hypothesis holds true because no significant difference could be found between the contexts for this condition as at any Ambisonic order; 64 kbps/ch Channel Mapping Family 3 showed a significant difference between contexts for third-order Ambisonic content only. For 32 kbps/ch Channel Mapping Family 2, the third null hypothesis is rejected because there were significant differences found between the contexts for this condition at all Ambisonic orders; 32 kbps/ch Channel Mapping Family 3 showed a significant difference between contexts for fifth-order Ambisonic content only. Finally, for 16 kbps/ch Channel Mapping Family 2, the third null hypothesis is rejected because there were significant differences found between the contexts for this condition at all Ambisonic orders; 16 kbps/ch Channel Mapping Family 3 showed a significant difference between contexts for fifth-order Ambisonic content only. In general, Channel Mapping Family 3 showed the least variation in median BAQ between contexts at most bitrates and Ambisonic orders.

3 DISCUSSION

A similar study, conducted by Fela et al. [10], involved the evaluation of 360° video with fourth-order Ambisonic audio reproduced over a loudspeaker array. The study also featured varied content with different contexts. However, all of the scenes in that study were real recordings, not virtually produced, and different orders of Ambisonics were

Table 3. The p values obtained using Kruskal-Wallis test on data grouped by different contexts at separate conditions and Ambisonic orders. Assuming a p-value threshold of 0.05, the significantly different rating score distributions between the contexts are marked by p-values in bold.

	low _anchor	mid _anchor	cm2 _16kbpspch	cm3 _16kbpspch	cm2 _32kbpspch	cm3 _32kbpspch	cm2 _64kbpspch	cm3 _64kbpspch	hid _reference
1OA	0.883807	**0.000002**	**0.000253**	0.450940	**0.007015**	0.081333	0.413270	0.503503	0.244155
3OA	0.217983	**0.000001**	**0.000704**	0.203719	**0.018235**	0.085015	0.222632	**0.008767**	0.066835
5OA	**0.001135**	**0.000002**	**0.000840**	**0.008950**	**0.002021**	**0.000002**	0.937564	0.663838	0.532762

not tested. Their study assessed video and audio separately, then assessed video and audio combined, whereas this study only focused on audio impairments; video remained constant in a given trial and was not assessed independently. This is a factor that shall be discussed further in SEC. 4. The Ambisonic audio in Fela et al.'s study was compressed at the same compression rates of 16, 32, and 64 kbps/ch using FFmpeg with Advanced Audio Codec–Low Complexity encoder, whereas this study used Opus audio codec with its channel mapping 2 and 3 families.

One finding in this study was that audio compressed at 64 kbps/ch was not significantly different than uncompressed audio in most cases. This could suggest that Opus preserves audio quality at this bitrate better than the codecs used in Fela et al.'s study, in which there was a more obvious reduction in quality when audio was compressed at 64 kbps/ch from the original pulse-code–modulation signal. However, previous evaluation of Opus-compressed Ambisonic scenes using loudspeaker reproduction [4] revealed significant differences between ratings of simple scenes (consisting of a single sound source) compressed at 64 kbps/ch and hidden reference at fifth-order Ambisonics. Accordingly, these findings might also suggest that Ambisonic audio rendering over loudspeakers leads to better discrimination of quality impairment by assessors rather than binaural reproduction using a standard virtual loudspeaker-based rendering and generic HRTFs. Finally, Fela et al. used a different metric to assess stimuli in their tests, Mean Opinion Score, whereas this study used BAQ as the metric to assess the stimuli.

4 FURTHER WORK

Because of a large number of variables in this study, data analysis was quite challenging, and many different approaches could have been taken. In further work, it would be useful to focus on a certain context at a certain Ambisonic order and give listeners fewer stimuli to compare at once, for example, cutting down from the current nine stimuli in each trial and instead having just five stimuli: one hidden reference, both anchors, and Channel Mapping Families 2 and 3 for just one bitrate of compression. This may help to reduce the spread exhibited in the results gathered in this study, which could give rise to a significant difference between the two different channel mapping families.

The trials presented could also be tested with no visuals. If there is a significant difference, such as a higher rating/less difference between compressed audio and the hidden reference with visuals enabled, this could suggest that lower bitrates can be used when the VR media contains visuals. Conversely, the perceived effects due to the increased complexity of the visuals could also be investigated, such as complex animations and more realistic assets.

Further exploration could also involve changing the content; for example, using real 360° videos and simultaneously recorded Ambisonic audio of real-life scenes. A comparison could then be made on the effects of Opus compression on real or virtual scenes and about whether one favors a certain channel mapping family over another; the only context in which this would be challenging is in "Game" scenes, because most VR games are created with virtual content. Using a microphone array capable of capturing fifth-order (or higher) Ambisonics would be the preferred recording option over up-mixing lower-order Ambisonic content.

Further work could look to improve the test procedure. For example, conducting a pilot test for the pre-screening of each participant as described in [5] could reduce the number of participants that had to be excluded from the final results. Also, a more sophisticated way to keep the visuals and audio in sync is desirable, in order for participants to be able to play certain parts of the scene and set a playback loop. In this study, the participants had to listen to the stimulus from the beginning when they switched conditions.

This study could also be extended to investigate other binaural rendering chains, for example, different HRTF preprocessing methods and different HRTFs.

5 CONCLUSION

This study has attempted to contribute new information about which bitrates of Opus compression and channel mapping families provide perceptual transparency when dealing with Ambisonic audio for different VR contexts and Ambisonic orders. The four contexts presented were gaming, music, soundscapes, and teleconferencing. The Opus parameters investigated were the channel mapping family and bitrate of compression.

The first null hypothesis investigated was "there is no significant difference between the BAQ rating of compressed and uncompressed stimuli." This null hypothesis holds true for 64 kbps/ch because no significant difference was found between this bitrate and the uncompressed hidden reference. For 32 kbps/ch, there were trials in which the compressed stimuli garnered a median BAQ significantly dif-

ferent than the uncompressed audio, and therefore, this null hypothesis is disproven in most of the trials at this bitrate. For 16 kbps/ch, in most trials, the compressed stimuli garnered a median BAQ significantly different than the hidden reference, therefore also disproving this null hypothesis at this bitrate.

The second null hypothesis was "the channel mapping family, within a certain bitrate of compression, has no effect on the BAQ rating that a stimulus is awarded." This null hypothesis holds true for 64 kbps/ch because there was no significant difference between the channel mapping families at any Ambisonic order. For 32 kbps/ch, there were significant differences between the channel mapping families in first-order and third-order Ambisonic content, which disproves this null hypothesis for this bitrate at these Ambisonic orders. For 16 kbps/ch, there were significant differences between the channel mapping families in third-order Ambisonic content, which disproves this null hypothesis for this bitrate at this Ambisonic order.

The third null hypothesis investigated was "for each of the codec parameters, the BAQ rating is independent of the scene context." This null hypothesis holds true for some bitrates and some Ambisonic orders but is disproven by others; in general, 64 kbps/ch Channel Mapping Family 2 garnered the lowest variation in median BAQ between different contexts, because there was no significant difference between context at any Ambisonic order—this does not necessarily mean that this condition produced the highest BAQ for each, just that it gained the most consistent BAQ rating across all contexts.

The key result is that across all trials, there was no significant difference between stimuli compressed at 64 kbps/ch, using Channel Mapping Family 3, and the hidden reference, making these settings optimal to use if the BAQ of the original Ambisonic audio is to be preserved. Channel Mapping Family 2 at this bitrate performed just slightly worse, garnering a BAQ rating significantly different than the uncompressed audio for only one of the scenes. Furthermore, Channel Mapping Family 3 showed no significant difference in median BAQ ratings across evaluated contexts at a higher number of bitrate and Ambisonic order conditions than Channel Mapping Family 2, which suggests it is a more robust compression scheme.

6 ACKNOWLEDGMENT

The authors gratefully acknowledge the participation of the listening test subjects. This research is supported by Google.

7 REFERENCES

[1] M. O'Neill, "How Netflix Bankrupted and Destroyed Blockbuster [INFOGRAPHIC]," *Business Insider* (2011 Mar.). https://www.businessinsider.com/how-netflix-bankrupted-and-destroyed-blockbuster-infographic-2011-3.

[2] M. Narbutt, S. O'Leary, A. Allen, J. Skoglund, and A. Hines, "Streaming VR for Immersion: Quality Aspects of Compressed Spatial Audio," presented at the *23rd International Conference on Virtual System Multimedia (VSMM)*, pp. 1–6 (Dublin, Ireland) (2017 Oct.).

[3] M. Narbutt, J. Skoglund, A. Allen, et al., "AMBIQUAL: Towards a Quality Metric for Headphone Rendered Compressed Ambisonic Spatial Audio," *Appl. Sci.*, vol. 10, no. 9, paper 3188 (2020 May). https://doi.org/10.3390/app10093188.

[4] T. Rudzki, I. Gomez-Lanzaco, P. Hening, et al., "Perceptual Evaluation of Bitrate Compressed Ambisonic Scenes in Loudspeaker Based Reproduction," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 Mar.), paper 100.

[5] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

[6] Sony Interactive Entertainment Europe Limited, "PlayStation VR," https://www.playstation.com/en-gb/ps-vr/ (accessed Sep. 24, 2021).

[7] Meta, "Oculus Quest 2: Our Most Advanced New All-in-One VR Headset," https://www.oculus.com/quest-2/ (accessed Sep. 7, 2021).

[8] E. Deleflie, "Interview With Simon Goodwin of Codemasters on the PS3 Game Dirt and Ambisonics," https://ambisonicbootlegs.wordpress.com/2007/08/30/interview-with-simon-goodwin-of-codemasters-on-the-ps3-game-dirt-and-ambisonics/ (2007 Sep).

[9] Apple, "Apple Music Announces Spatial Audio With Dolby Atmos and Lossless Audio," https://www.apple.com/uk/newsroom/2021/05/apple-music-announces-spatial-audio-and-lossless-audio/ (2021 May).

[10] R. F. Fela, A. Pastor, P. L. Callet, et al., "Perceptual Evaluation on Audio-Visual Dataset of 360 Content," *arXiv preprint arXiv:2205.08007* (2022 May).

[11] D. I. Johnston, H. W. Egermann, and G. C. Kearney, "Measuring the Behavioral Response to Spatial Audio Within a Multi-Modal Virtual Reality Environment in Children With Autism Spectrum Disorder," *Appl. Sci.*, vol. 9, no. 15, paper 3152 (2019 Aug.). https://doi.org/10.3390/app9153152.

[12] M. Wang and E. Anagnostou, "Virtual Reality as Treatment Tool for Children With Autism," in V. B. Patel, V. R. Preedy, and C. R. Martin (Eds.), *Comprehensive Guide to Autism*, pp. 2125–2141 (Springer, New York, NY, 2014).

[13] R. Fino, M. J. Lin, A. Caballero, and F. F. Balahadia, "Disaster Awareness Simulation for Children With Autism Spectrum Disorder Using Android Virtual Reality," *J. Telecommun. Electron. Comput. Eng.*, vol. 9, no. 2-6, pp. 59–62 (2017 Jun.).

[14] M. L.and J. C. Burke Max, "Virtual Reality for Autism Communication and Education, With Lessons for Medical Training Simulators," in K. S. Morgan, H. M. Hoffman, D. Stredney, and S. J. Weghorst (Eds.), *Medicine Meets Virtual Reality*, Studies in Health Technology and Informatics, vol. 39, pp. 46–53 (IOS Press, Amsterdam, The Netherlands, 1997).

[15] K. A. Karl, J. V. Peluchette, and N. Aghakhani, "Virtual Work Meetings During the COVID-19 Pandemic:

The Good, Bad, and Ugly," *Small Group Res.*, vol. 53, no. 3, pp. 343–365 (2022 Jun.).

[16] B. Lee, "Test Stimuli for Context Based Evaluation of the OPUS Audio Codec," *Zenodo* (2022 Jul.). http://doi.org/10.5281/zenodo.6906836.

[17] Cycling '74, "What is Max?" https://cycling74.com/products/max (accessed Sep. 23, 2021).

[18] D. Rudrich, "IEM Plug-in Suite," https://plugins.iem.at/ (accessed Sep. 23, 2021).

[19] M. Kronlachner, "ambiX v0.2.8 – Ambisonic Plug-In Suite," http://www.matthiaskronlachner.com/?p=2015 (accessed Aug. 10, 2019).

[20] Spatial Audio for Domestic Interactive Entertainment, " SADIE II Database," https://www.york.ac.uk/sadie-project/database.html (accessed August 10, 2019).

[21] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, paper 2029 (2018 Oct.).

[22] J. Skoglund and M. Graczyk, "Ambisonics in an Ogg Opus Container," *RFC 8486* (2018 Oct.). https://tools.ietf.org/html/rfc8486.html.

[23] T. Rudzki, C. Earnshaw, D. Murphy, and G. Kearney, "SALTE Pt. 2: On the Design of the SALTE Audio Rendering Engine for Spatial Audio Listening Tests in VR," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), e-Brief 537.

[24] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Berry, and A. Garcia, "A Fifty-Node Lebedev Grid And Its Applications To Ambisonics," *J. Audio Eng. Soc.*, vol. 64, no. 11, pp. 868–881 (2016 Dec.). http://www.aes.org/e-lib/browse.cfm?elib=18524.

[25] SideQuest, "SideQuest: Early Access Virtual Reality," https://sidequestvr.com/ (accessed Aug. 11, 2021).

[26] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," *Am. Stat.*, vol. 32, no. 1, pp. 12–16 (1978 Feb.).

THE AUTHORS


Ben Lee


Tomasz Rudzki


Jan Skoglund


Gavin Kearney

Ben Lee graduated from University College London in 2018 with a B.Sc. in Physics and undertook an M.Sc. in Audio and Music Technology at the University of York. His M.Sc. project focused on the perceptual evaluation the Opus audio codec. Upon obtaining his M.Sc., he worked for a year in acoustic consultancy but later returned to York's AudioLab as a research technician as part of a team at the AudioLab, which focuses on many different applications of immersive and interactive audio. In his spare time, he enjoys electronic music production or spending his time outdoors, hiking, or cycling.

•

Tomasz Rudzki received his B.Sc. and M.Sc. degrees in Telecommunications from the Warsaw University of Technology in 2009 and 2014, respectively. He is currently pursuing a Ph.D. degree in Electronic Engineering at the University of York. In 2021, he worked as a Research Intern within the Meta Reality Labs Research Audio Team. His research interests include signal processing, auditory perception, perceptual evaluation of spatial audio codecs, indirect evaluation methods for spatial audio, and HRTF personalization.

•

Jan Skoglund leads a team at Google in San Francisco, CA, developing speech and audio signal processing components for capture, real-time communication, storage, and rendering. After receiving his Ph.D. degree at Chalmers University of Technology in Sweden, 1998, he worked on low bitrate speech coding at AT&T Labs-Research, Florham Park, NJ. He was with Global IP Solutions (GIPS), San Francisco, CA, from 2000 to 2011 working on speech and audio processing, such as compression, enhancement, and echo cancellation, tailored for packet-switched networks. GIPS's audio and video technology was found in many deployments by, e.g., IBM, Google, Yahoo, WebEx, Skype, and Samsung, and was open-sourced as WebRTC after a 2011 acquisition by Google. Since then, he has been a part of Chrome at Google.

•

Gavin Kearney is a Professor of Audio Engineering at the University of York. He graduated from Dublin Institute of Technology in 2002 with an Honors degree in Electronic Engineering and has since obtained M.Sc. and Ph.D. degrees in Audio Signal Processing from Trinity College Dublin. He joined the University of York in 2011. He has written over 100 research articles and patents on different aspects of immersive and interactive audio and leads a team of researchers in this area at the AudioLab at York. He is currently Vice-Chair of the AES Audio for Games Technical Committee and an active sound engineer and producer of immersive audio experiences.