2017-10

Streaming VR for Immersion: Quality Aspects of Compressed Spatial Audio

Miroslaw Narbutt
Technological University Dublin, miroslaw.narbutt@tudublin.ie

Sean O'Leary
Technological University Dublin, sean.oleary@tudublin.ie

Andrew Allen
Google, Inc., bitllama@google.com

See next page for additional authors

Recommended Citation

Narbutt, M., O'Leary, S. & Allen, A. (2017). Streaming VR for Immersion: Quality aspects of Compressed Spatial Audio. *23rd International Conference on Virtual Systems and Multimedia (VSMM2017)*, Dublin, Belfast, October 2017. doi:10.1109/VSMM.2017.8346301

Funder: Google, Inc.

Streaming VR for Immersion: Quality aspects of Compressed Spatial Audio

Miroslaw Narbutt
School of Computing
Dublin Institute of Technology
Dublin, Ireland
miroslaw.narbutt@dit.ie

Seán O'Leary
School of Computing
Dublin Institute of Technology
Dublin, Ireland
sean.oleary@dit.ie

Andrew Allen
Google, Inc.
San Francisco, Ca., U.S.A.
bitllama@google.com

Jan Skoglund
Google, Inc.
San Francisco, Ca., U.S.A.
jks@google.com

Andrew Hines
School of Computer Science
University College Dublin
Dublin, Ireland
andrew.hines@ucd.ie

***Abstract*—Delivering a 360-degree soundscape that matches full sphere visuals is an essential aspect of immersive VR. Ambisonics is a full sphere surround sound technique that takes into account the azimuth and elevation of sound sources, portraying source location above and below as well as around the horizontal plane of the listener. In contrast to channel-based methods, ambisonics representation offers the advantage of being independent of a specific loudspeaker set-up.**

Streaming ambisonics over networks requires efficient encoding techniques that compress the raw audio content without compromising quality of experience (QoE). This work investigates the effect of audio channel compression via the OPUS 1.2 codec on the quality of spatial audio as perceived by listeners. In particular we evaluate the listening quality and localization accuracy of first-order ambisonic audio (FOA) and third-order ambisonic audio (HOA) compressed at various bitrates (i.e. 32, 64, 128 and 128, 256, 512kbps respectively).

To assess the impact of OPUS compression on spatial audio a number of subjective listening tests were carried out. The sample set for the tests comprises both recorded and synthetic audio clips with a wide range of time-frequency characteristics. In order to evaluate localization accuracy of compressed audio a number of fixed and dynamic (moving vertically and horizontally) source positions were selected for the test samples. The results show that for compressed spatial audio, perceived quality and localization accuracy are influenced more by compression scheme, bitrate and ambisonic order than by sample content. The insights provided by this work into factors and parameters influencing QoE will guide future development of a objective spatial audio quality metric.

***Keywords*—virtual reality, spatial audio, ambisonics, audio coding, audio compression, opus codec, MUSHRA**

I. Introduction

Research and development for virtual and augmented reality systems has led to a growing number of consumer-grade virtual reality (VR) headsets. The quality of experience (QoE) [1] for users can be quantified from a variety of different perspectives, e.g. fidelity, immersion, and presence. Gilbert [2] terms this "authenticity", i.e. a function of how the virtual environment provides the conscious and unconscious experience expected by the user. Depending on the application, senses are prioritised or weighted differently. Some VR experts speculate that in VR production, audio is the most important component of believable immersion [3]. It has been shown that including white noise in a scene that otherwise has no virtual audio improves the sense of presence in the scene [4]. A 360-degree soundscape that matches the 3D visual is important for sustained immersion. Immersion relies heavily on how audio is propagated within the scene. Audio cues must occur in the right direction and plane, with the right intensity and need to be rendered in real-time to match head movements. Gilbert argues that fidelity offers no real guidelines for establishing target thresholds: visual fidelity may be important for a vehicle simulator but for heavy industrial equipment where decisions are made based on engine sounds, audio fidelity may be preferred. Whether this audio fidelity requires timbral quality or localization accuracy is another consideration. For instance, faithful reproduction without compression artefacts is a priority for an orchestral recital while the location of a sound may be more critical is an action game environment.

Advances in microphone arrays and spatial technologies for capture and reproduction [5] have increased the opportunity to integrate soundscapes into VR mediums. A variety of application from animated movies and games through to applications such as urban planning are availing of spatial audio in virtual and augmented reality implementations [6]. Consumer VR headsets are designed for personalised immersive media consumption. Combined with headphones, an immersive experience can be rendered in realtime using spatially captured, encoded and transported sound to provide a binaural experience that matches head orientation and movement. Capture, encoding and transmission of spatial audio content to deliver an immersive auditory experience can be handled in a variety of ways. Sound field synthesis solutions such as

This publication has emanated from research conducted in the CONNECT research centre in collaboration with Google, Inc. with the financial support of Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

ambisonics have bandwidth efficiencies over pure object based solutions [7]. The methods used in encoding will impact the fidelity of the media content created. An understanding of the trade-offs between compression for bandwidth management and quality of experience will inform the impact of current limited bandwidth realities on the longer term digital heritage in terms of compression and media legacy [8].

This paper presents the results of experiments that explore the relationship between audio quality and localisation accuracy for uncompressed audio and OPUS-compressed audio samples. The results will be used to validate and optimise work currently underway to develop a full reference objective spatial audio quality metric adapted from ViSQOLAudio [9], [10].

II. Background

A. Ambisonics

Ambisonics is an approach to encoding and rendering spatialised audio, initially developed in the 1970s. In contrast to channel-based methods, ambisonics representation offers the advantage of being independent of a specific loudspeaker set-up. It may be rendered to set-ups consisting of only few loudspeakers or even to headphones using binaural rendering.

Ambisonics is a full-sphere surround sound technique that takes into account the azimuth and elevation of sound sources, above and below as well as around the listener. The signal for a given sound source can be represented as a sound field using a spherical decomposition with the B-format standard and scaled to any desired spatial resolution. For example, First-order Ambisonics (FOA) audio is encoded into four channels: an omnidirectional gain and three dimensional components: forward/backwards, left/right, and up/down [11]. Moving to Higher Order Ambisonics (HOA) significantly improves the Quality of Experience (QoE). The downside to ambisonics B-format is the large amount of data and processing power required by HOA to transform a collection of multichannel signals into a rendered soundscape e.g. third-order ambisonics requires a 16 channel B-format.

B. OPUS 1.2 codec with Channel Mapping Family 2

The Opus audio codec [12] was ratified by the Internet Engineering Task Force (IETF) in 2012 as an audio streaming and storage standard. It has gained popularity through adoption by companies such as Skype and Google and its open and royalty-free licensing conditions. It is ubiquitous is both browsers and mobile apps having been included in real-time WebRTC API and YouTube streaming service [7]. This work investigates the impact of encoding ambisonic B-format content using the OPUS 1.2 codec with Channel Mapping Family 2 implementation [13]. The quality of third-order and first-order ambisonic audio at a variety of bitrates are evaluated.

C. Aims

Streaming ambisonic data over networks requires efficient encoding techniques that compress the raw audio content in real-time and without significantly compromising QoE. To develop a suitable approach to compression of ambisonic B-format its performance must be evaluated using formalised quality judgement experiments [7]. In the absence of objective assessment methods, this is done usually with a panel of experienced listeners who evaluate listening audio quality [14].

III. Method

Listening tests were carried out using the MUSHRA test methodology (MUltiple Stimuli with Hidden Reference and Anchor) following the ITU-R BS.1534-3 recommendation [15]. During a MUSHRA test, listeners are presented with a labelled reference and a number of unlabelled test samples (stimuli or condition). Their task is to assign ratings to the unlabelled samples using a numerical continuous scale ranging from 0 to 100 in five descriptive intervals: bad (0-20), poor (20-40), fair (40-60), good (60-80), and excellent (80-100). The assessor can switch at will between the reference audio and any of the stimuli under test. The MUSHRA test procedure is typically used to evaluate intermediate levels of audio impairment. It is a double-blind multi-stimulus test method; one unaltered version of the reference and one or more anchor samples are hidden amongst multiple test conditions.

A. Test Setup

For the tests three audio samples (of 7 to 15 seconds duration) covering a variety of musical sounds were selected from CDs and the EBU music database [16]. These recorded audio clips have been chosen as particularly difficult to compress [17]. An additional six audio samples were synthetically generated to cover a wide range of time-frequency characteristics (i.e. chirp signal and various kinds of colored noise). Details of the samples can be found in Table 1.

TABLE I
AUDIO SAMPLES USED DURING PILOT TESTS

Label	Music Type	Source
Vega	Vocals (Suzanne Vega)	CD
Castanets	Castanets	EBU
Glock	Glockenspiel	EBU
Chirp	Synthetic chirp signal	synthetic
FreqBands	Noise from 32 consecutive Frequency Bands	synthetic
PinkNoise	Bursty Pink Noise	synthetic
WhiteNoise	Bursty White Noise	synthetic
BlueNoise	Bursty Blue Noise	synthetic
VioletNoise	Bursty Violet Noise	synthetic

All clips had a sampling frequency of 48 kHz (one was resampled from 44.1kHz) and were recorded in stereo format. These were then converted to mono format, encoded to first-order (FOA) and third-order (HOA) ambisonic audio with a variety of localizations as shown in Figure 1.

Ambisonic audio signals were encoded to a variety of bitrates to produce a range of conditions, and finally rendered to a binaural format for presentation. Each rendered audio signal (i.e. treatment) was evaluated with 7 conditions, namely, Reference, HOA512, HOA256, HOA128, FOA128, FOA64,

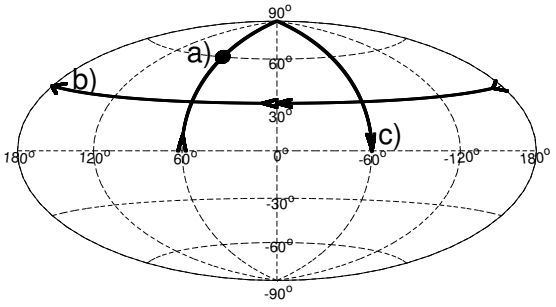

Fig. 1. Localization of sound sources: a) fixed localization (azimuth 60°, elevation 60°, b) dynamic azimuth localization with audio source moving horizontally (i.e. rotating azimuth above the listener's head), c) dynamic elevation localization with audio source moving vertically (i.e. moving up in elevation on the left hand side, then down on the right hand side).

and anchor. Condition Reference was the original, uncompressed audio, which also served as the hidden reference. Conditions HOA512, HOA256, and HOA128 represent 3rd-order ambisonics audio (i.e. HOA) encoded with 512, 256, and 128kbps respectively. Conditions FOA128, FOA64 represent first-order ambisonic audio (i.e. FOA) encoded at 128 and 64kbps respectively. Finally, condition "anchor" represents first-order ambisonic audio encoded at 32kbps and served as the hidden anchor of this MUSHRA test. Details of encoding schemes and bitrates used during pilot tests can be found in Table 2.

TABLE II
ENCODING/COMPRESSION SCHEMES USED DURING PILOT TESTS

Type	ambisonics order	bitrate (kbps)	bitrate per channel (kbps)
Reference	3	12288	768
HOA 512	3	512	32
HOA 256	3	256	16
HOA 128	3	128	8
FOA 128	1	128	32
FOA 64	1	64	16
FOA 32 (anchor)	1	32	8

B. Testing platform and testing environment

The open-source WebMUSHRA testing platform [18] which implements ITU-R BS.1534 recommendation was selected for the listening tests. It is written as an HTML5 web application (using JavaScript and WebAudio API) and can be run within a web browser.

Figure 2 shows the Graphical User Interface which was used for the listening tests. Under each "Play" button, with the exception of the button for the reference, a slider is used to grade the quality of the test items according to the continuous quality scale used. For each of the test items, the order of the conditions being tested is randomised to avoid sequential effects.

Most of the listening tests took place in a controlled environment, i.e. recording studios at Trinity College Dublin

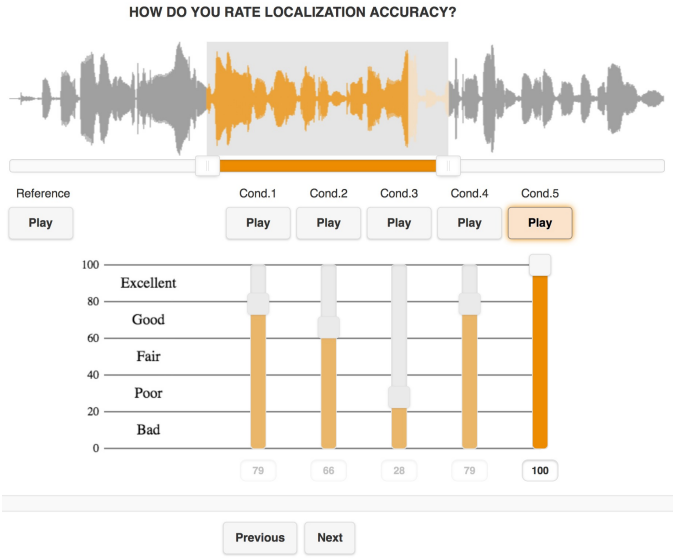

Fig. 2. WebMUSHRA Graphical User Interface captured on MacBook Pro web browser

and Dublin Institute of Technology. The WebMUSHRA test system was presented to assessors using a laptop (MacBook Pro). Professional monitoring headphones were used for the majority of the tests (i.e. Audio-Technica ATH-M70x with flat frequency response from 5Hz to 40kHz). The average time taken for the listening test (including training phase) was 50 minutes.

C. Pilot tests

Two pilot tests were performed before the actual tests to establish test duration, test questions (how to assess localization), sample content, and sample localizations. As a result, seven audio samples were chosen for the actual listening tests. Details of these samples can be found in Table 3.

TABLE III
AUDIO SAMPLES USED DURING LISTENING TESTS

Label	Music Type	Source
Vega	Vocals (Suzanne Vega)	CD
Castanets	Castanets	EBU
Glock	Glockenspiel	EBU
VegaRev	Vocals (Suzanne Vega) w. Reverb Effect	processed CD
CastanetsRev	Castanets w. Reverb Effect	processed EBU
PinkNoiseRev	Bursty Pink Noise w. Reverb Effect	synthetic

Details of encoding schemes and bitrates used during listening tests can be found in Table 4.

D. Selection of assessors

21 participants (20 male and 1 female) performed the test, with ages ranging from 20 to 53 years old, and an average age of 32. The participants included experienced listeners (9 subjects) comprising of professional audio engineers, and

TABLE IV
ENCODING/COMPRESSION SCHEMES USED DURING LISTENING TESTS

Type	ambisonics order	bitrate (kbps)	bitrate per channel (kbps)
Reference	3	12288	768
HOA 512	3	512	32
HOA 256	3	256	16
FOA 128	1	128	32
FOA 32 (anchor)	1	32	8

academics with prior experience of similar tests, and semi-experienced listeners (12 subjects) comprising of post-graduate students in Trinity College Dublin and Dublin Institute of Technology. Experiments followed a methodology approved by Dublin Institute of Technology's Ethics Committee.

E. Training phase

It was deemed mandatory to train the assessors at special training sessions in advance of the test to obtain reliable results. Accordingly, before the actual listening tests, the assessors had to complete a training phase. This training phase was to allow the assessors some time to familiarize themselves with the Graphical User Interface, testing procedures and the evaluation method which included two metrics:

Listening Quality - the perceived similarity in sound quality of the test samples compared to the reference;

Localization Accuracy - the perceived similarity of the location of the sound sources in the test samples when compared to the reference.

F. Grading phase

Audio listening tests were carried out using the MUSHRA test methodology. During the listening tests, each listener was asked to evaluate 9 different sample sets (3 audio clips with fixed localizations, 3 with variable Azimuth, and 3 with variable Elevation angle), each sample set consisting of 5 conditions table IV. Listeners were asked to rate the stimuli in regard to two aspects: listening quality and localization accuracy. This makes total 18 test pages per listening test.

IV. RESULTS

The MUSHRA test procedure recommends excluding subjects who give the hidden reference a score of less than 90 for more than 15% of the test items. Accordingly, the results of 3 participants were therefore excluded from the results.

Aggregated MUSHRA scores by encoding schemes for both Listening Quality and Localization Accuracy with 95% confidence intervals are shown in Figures 3 and 4 respectively.

A. MUSHRA scores by encoding scheme

Figure 3 depicts the aggregated mean values of the Listening Quality scores and the 95% confidence intervals for five encoding schemes (i.e. REF, HOA512, HOA256, FOA128, FOA32). The aggregated quality scores are shown here as the average MUSHRA score obtained for all nine audio test samples. All are significant ($p <= 0.0001$) except for

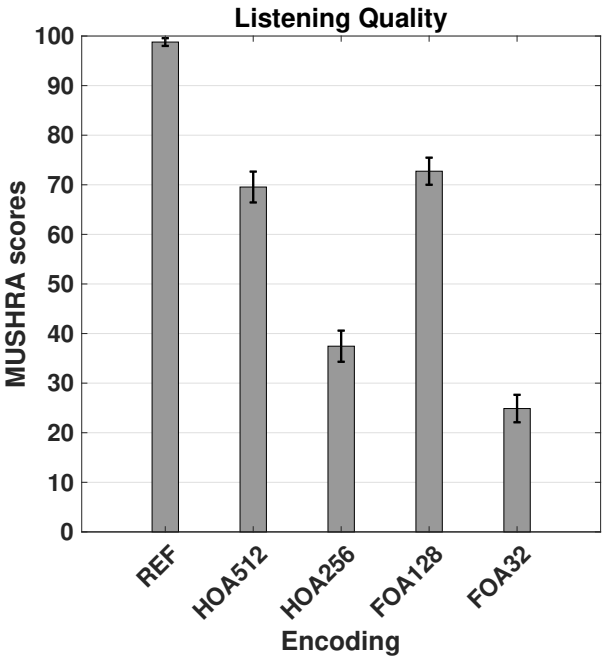

Fig. 3. Listening Quality MUSHRA scores aggregated by encoding scheme. Mean values with 95% confidence intervals shown.

the HOA512 vs FOA128 (p=0.45). It can be observed that both HOA512 and FOA128 encoding schemes (both with bitrates of 32kbps per channel) perform well and achieve "good" listening quality. Also, the aggregated mean values and 95% confidence intervals are comparable for both HOA512 and FOA128 with FOA128 performing slightly better. The HOA256 and FOA32 samples (with bitrates of 16kbps and 8kbps per channel respectively) perform poorly, both achieving and average rating of less than 40 on the MUSHRA scale.

Figure 4 presents the aggregated mean values of the Localization Accuracy scores and the 95% confidence intervals for 5 encoding schemes. As before, the aggregated Localization Accuarcy scores are taken to be the average MUSHRA scores obtained for all nine audio test samples with $p <= 0.0001$. It can clearly be seen that HOA512 outperforms FOA128 encoding scheme in regard to localization accuracy (i.e. excellent/good), confirming that under the tested encoding scheme 512kbps Higher Order Ambisonics significantly improves the Quality of Experience. For localization accuracy the FOA128 encoding scheme outperforms the HOA256 (i.e. good/fair). This is likely due to the difference in bitrates per channel (32kbps and 16kbps respectively).

The listening test results were re-analysed in more detail to investigate differences between MUSHRA scores for a given sample's content. Figure 5 (Listening Quality) and Figure 6 (Localization Accuracy) show a breakdown of the results by encoding scheme per each of the nine audio samples. It can be observed that the average MUSHRA scores obtained for HOA vary more than for FOA encoding schemes. This is true for both Listening Quality and Localization Accuracy

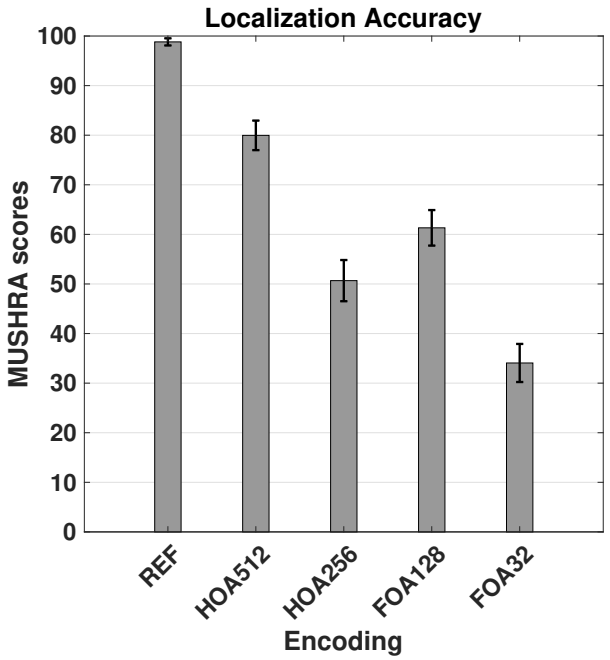

Fig. 4. Localization Accuracy MUSHRA scores aggregated by encoding scheme. Mean values with 95% confidence intervals shown.

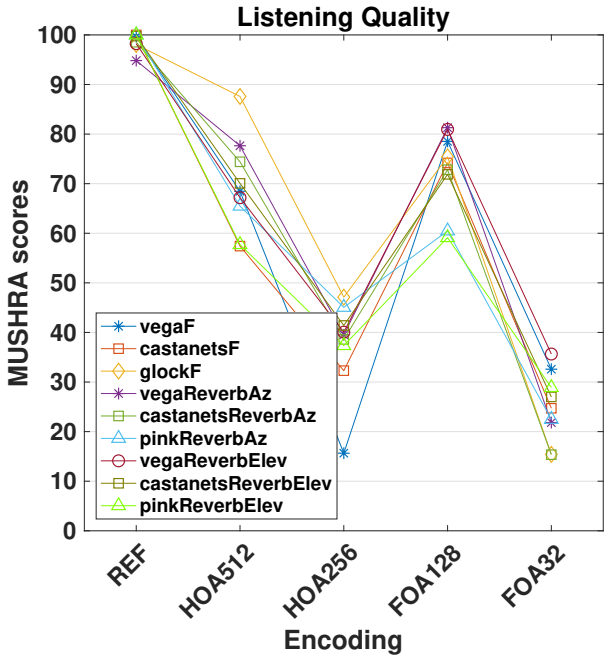

Fig. 5. Listening Quality MUSHRA scores by encoding scheme (broken down by sample type)

assessments.

B. Mushra scores by content

Aggregated MUSHRA scores by content for both Listening Quality and Localization Accuracy are shown in Figures 7 and 8 respectively.

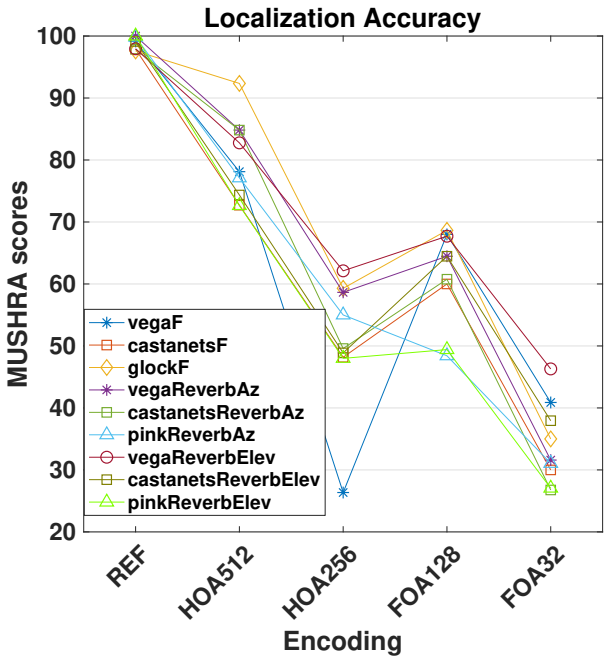

Fig. 6. Localization Accuracy MUSHRA scores by encoding scheme (broken down by sample type)

It can be observed that the MUSHRA scores obtained during listening tests did vary with content. Different test samples were assigned different MUSHRA scores for a given encoding scheme. This was more apparent in HOA512 than FOA128 encoding. However, the variation in quality was not consistent for a given test sample across all encoding schemes.

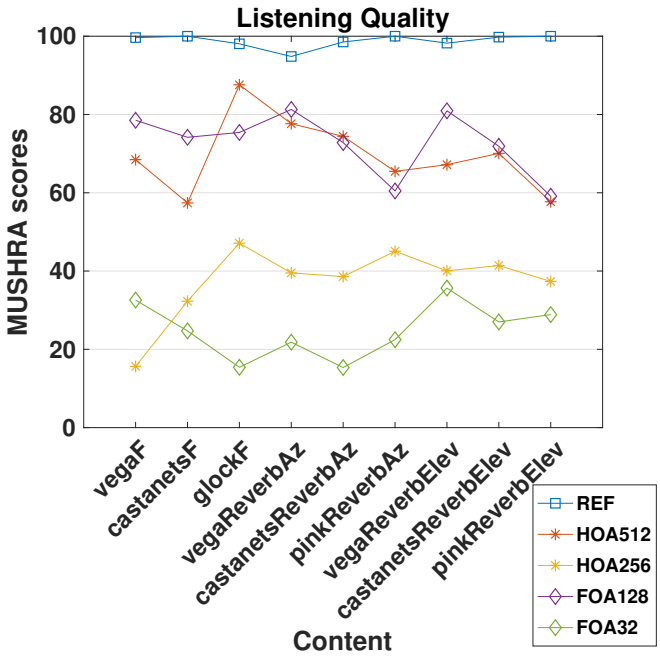

Fig. 7. Listening Quality MUSHRA scores by content.

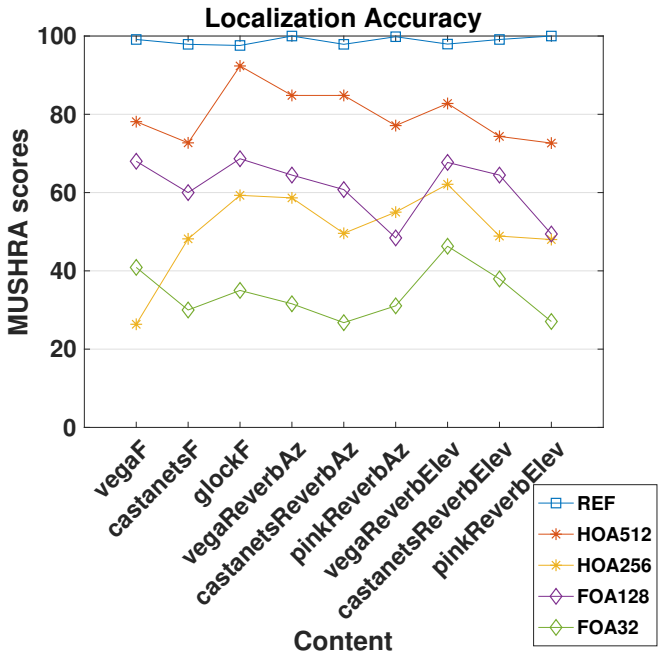

Fig. 8. Localization Accuracy MUSHRA scores by content.

V. Discussion

Audio listening tests confirmed that audio channel compression via the OPUS 1.2 codec can greatly impact both Listening Quality and Localization Accuracy. Bitrate per channel is key. Both HOA512 and FOA128 encoding schemes (32kbps per channel) perform "good" in regard to Listening Quality and HOA512 performs "excellent" in regards to Localization Accuracy. However, for the lower bitrates per channel tested, the OPUS codec (ver 1.2) had a detrimental effect on both Listening Quality and Localization Accuracy. For example, it has been observed that with bitrates of 16kbps per channel (used by HOA256 encoding scheme) HOA no longer outperforms FOA.

VI. Conclusions and Future Work

We investigated via audio listening tests the effect of audio channel compression (using OPUS codec ver1.2) on the quality of spatial audio as perceived by listeners. In particular we evaluated the listening quality and localization accuracy of first-order ambisonic audio (FOA) and third-order ambisonic audio (HOA) compressed at various bitrates. Results show that audio channel compression can greatly impact both Listening Quality and Localization Accuracy. These results will be used to validate and optimise work currently underway to develop a full reference objective spatial audio quality metric adapted from ViSQOLAudio [9], [10].

Acknowledgements

Thanks to Marcin Gorzel and Ian Kelly (Google, Inc.), Francis M. Boland and Enda Bates (Trinity College Dublin) for advice and input relating to spatial audio and ambisonic encoding quality.

References

[1] European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), "Qualinet White Paper on Definitions of Quality of Experience," March, 2013.

[2] S. B. Gilbert, "Perceived realism of virtual environments depends on authenticity," *Presence: Teleoperators and Virtual Environments*, vol. 25, no. 4, pp. 322–324, 2016.

[3] M. Kuhn, "Spatial audio and immersion," *Brown University*, March 2017.

[4] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the role of presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, 1997. [Online]. Available: http://dx.doi.org/10.1162/pres.1997.6.6.603

[5] E. Bates, S. Dooney, M. Gorzel, H. O'Dwyer, L. Ferguson, and F. M. Boland, "Comparing ambisonic microphones: Part 2," in *Audio Engineering Society Convention 142*, May 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18607

[6] J. Y. Hong, J. He, B. Lam, R. Gupta, and W.-S. Gan, "Spatial audio for soundscape design: Recording and reproduction," *Applied Sciences*, vol. 7, no. 6, 2017.

[7] J. Brettle and J. Skoglund, "Open-source spatial audio compression for vr content," in *SMPTE 2016 Annual Technical Conference and Exhibition*, Oct 2016, pp. 1–9.

[8] "Lossless compression and the future of memory," *Interactions: Studies in Communication Culture*, vol. 8, no. 1, 2017.

[9] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective assessment of perceptual audio quality using ViSQOLAudio," *IEEE Transactions on Broadcasting*, vol. PP, no. 99, pp. 1–13, 2017.

[10] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.

[11] A. Heller, R. Lee, and E. Benjamin, "Is my decoder ambisonic?" in *Audio Engineering Society Convention 125*, Oct 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14705

[12] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," *IETF*, September, 2012. [Online]. Available: RFC 6716, http://www.ietf.org/rfc/rfc6716.txt

[13] J. Skoglund and M. Graczyk, "IETF internet-draft: Ambisonics in an ogg opus container," 2017. [Online]. Available: http://tools.ietf.org/html/draft-ietf-codec-ambisonics-02

[14] ITU, "ITU-T methods for subjective determination of transmission quality," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.800, 1996.

[15] ITU, "ITU-R Rec. BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems," 2015.

[16] EBU Tech, "3253-E,Sound quality assessment material," *SQUAM CD (Handbook). EBU Technical Centre Brussels*, 1988.

[17] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "Perceived audio quality for streaming stereo music," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1173–1176.

[18] "A MUSHRA compliant web audio API based experiment software." [Online]. Available: https://github.com/audiolabs/webMUSHRA