Perceived Audio Quality for Streaming Stereo Music

Andrew Hines[†*]
andrew.hines@tcd.ie

Eoin Gillen[†]
ogiollae@tcd.ie

Damien Kelly[‡]
damien.kelly@google.com

Jan Skoglund[‡]
jks@google.com

Anil Kokaram[‡]
anil.kokaram@google.com

Naomi Harte[†]
nharte@tcd.ie

[†]Sigmedia, Dept. Electronic and Electrical Engineering, Trinity College Dublin, Ireland
[‡]Google, Inc., Mountain View, California 94043, USA

ABSTRACT

Users of audio-visual streaming services expect an ever increasing quality of experience. Channel bandwidth remains a bottleneck commonly addressed with lossy compression schemes for both the video and audio streams. Anecdotal evidence suggests a strongly perceived link between bit rate and quality. This paper presents three audio quality listening experiments using the ITU MUSHRA methodology to assess a number of audio codecs typically used by streaming services. They were assessed for a range of bit rates using three presentation modes: consumer and studio quality headphones and loudspeakers. Our results indicate that with consumer quality headphones, listeners were not differentiating between codecs with bit rates greater than 48 kb/s ($p>=0.228$). For studio quality headphones and loudspeakers aac-lc at 128 kb/s and higher was differentiated over other codecs ($p<=0.001$). The results provide insights into quality of experience that will guide future development of objective audio quality metrics.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and music Computing

Keywords

Audio Quality; MUSHRA; YouTube; Audio Codec

1. INTRODUCTION

With the rise of YouTube, Netflix, Hulu, Spotify, Pandora and others, it is clear that media streaming has become an established industry. One constant challenge is to deliver an acceptable quality of experience (QoE) [2] across the diverse range of devices on which content is consumed (e.g. mobile, desktop, home theatre) under variable network bandwidth constraints. To accommodate these diverse environments both the video and audio content are transmitted in compressed form using lossy compression schemes. Suitable objective metrics could help automate the evaluation of changes in QoE as a result of this transcoding process. As is well known, the development of these metrics must start with the collection of ground truth quality information. This paper considers two questions in the context of stereo music: i) can the listener perceive differences in the audio quality for the codecs tested and ii) how does presentation mode influence the results, i.e. is the listening hardware a dominant factor over the codec and bit rate? Anecdotal opinions on these issues are pervasive across many internet forums. Headphone listening quality [9], audibility of codec degradations in rooms [14], and codec audio quality [3] have been reported but limited laboratory testing has been published on the interaction between codec and presentation mode for a range of codecs. This paper reports on experiments that provide this ground truth information. A range of codecs and bit rates ("treatments") are examined: aac-he and aac-lc [8] codecs at four bit rates and examples of MP3 and OPUS [16] codecs.

2. MUSHRA AUDIO QUALITY TESTING

The MUSHRA standard test methodology is defined by ITU-R recommendation BS.1534.1 [13]. During a MUSHRA test, listeners are presented with a labeled reference and a number of unlabelled test samples (stimuli). The listener assigns 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). An unaltered version of the reference and one or more anchor samples are hidden amongst the treatments under test. The anchors are low-pass filtered versions of the reference. The methodology has been used in a variety of tests showing a good ability to rank low bit rate codecs [10, 7]. Other audio quality test methodologies exist, but for low bit rate codec testing, MUSHRA is a good compromise between an absolute category rating test (e.g. ITU-R BS.1284-1) and a test for almost undetectable impairments (e.g. ITU-R BS.1116-1). Biases in MUSRHA tests have been reported due to stimulus spacing and range equalising effects [18]. A HTML5 user interface for conducting MUSHRA tests via a tablet computer was developed fol-

*Supported by Google, Inc. and Enterprise Ireland, Innovation Partnership Project IP-2013 0232.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

ACM 978-1-4503-3063-3/14/11...$15.00.
http://dx.doi.org/10.1145/2647868.2655025.

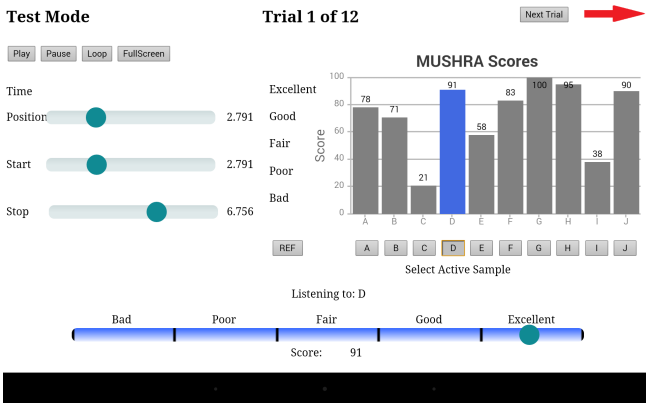

Figure 1: Example of MUSHRA UI screen captured on Nexus 7 tablet

lowing the guidelines set down in the recommendation [13] and is shared for use in research[1]. The layout was adapted to fit within a 7-inch tablet display and to allow touchscreen sliders. The design was altered to replace multiple sliders with a chart which keeps track of the scores assigned in the current trial. A single horizontal score slider below the chart adjusts the score of the currently-active treatment. Treatments are selected using alphabetized buttons below the chart. An example screen is shown in Figure 1. The user interface handles randomisation of the stimuli for presentation during testing. The samples are served from a local web server running on the tablet to remove latency in buffering the samples.

3. METHODS

Experiments were carried out using the MUSHRA test methodology for three presentation modes: loudspeakers and two types of headphones. Stereo music samples of 7 – 15 seconds duration were used covering a variety of musical sounds. Details of the samples can be found in Table 1. The files were all originally sampled at either 48 kHz or 44.1 kHz, 16 bit stereo (so for 44.1 kHz 2-channel audio, the bit rate is 1411.2 kb/s). Reference PCM WAV files were created at 48 kHz for all files. These were then transcoded and resampled using ffmpeg with Fraunhofer aac encoder for aac, libmp3lame for MP3 and libopus for Opus 1.1 to produce a range of treatments that were then formatted as WAV PCM files for presentation. The treatments used are presented in Table 1. There was no perceived level difference between the reference samples and the transcoded treatments. Stereo music clips were used, originating from CDs and the EBU music database [1]. All clips had a sampling frequency of 48 kHz. The first test used 10 samples which matched those used by Hoene et al.[7] in MUSHRA tests to evaluate the OPUS codec. The type of music in each sample is listed in Table 1. The second test kept six of the samples from the first test but replaced 4 samples with 6 alternative samples taken from the EBU database. The substituted samples were replaced with alternatives that are more sensitive to bit-rate reduction and frequency response as indicated in [1]. As in [7], two hidden anchors were used, one lowpass-filtered at 3.5 kHz and one lowpass-filtered at 7.0 kHz. Along with the hidden reference and two anchors, 7

[1]Download available at http://www.sigmedia.tv/tools

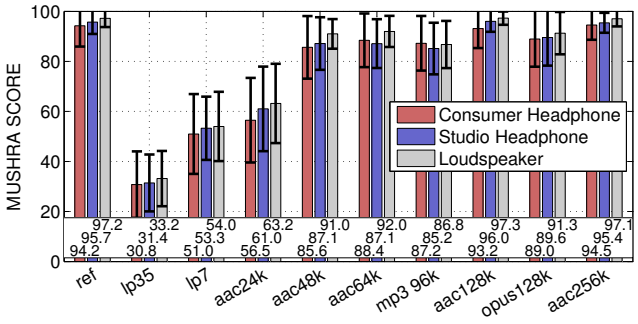

Figure 2: Results by treatment and presentation mode (experiments 1–3) for 6 music samples common to all experiments. Mean values with 95% confidence intervals shown

Table 1: Music Samples

Label	Music Type	Source	Exp.
Boz	Rock/R&B (Boz Scaggs)	CD	1,2,3
Steely	Soft Rock (Steely Dan)	CD	1,2,3
Champs	Rock (Queen)	CD	1
Harry	Jazz (Harry James)	CD	1
Purcell	Classical (Purcell)	CD	1
Electro	Electronica (Matmos)	CD	1
Castanets	Castanets	EBU	1,2,3
Moonlight	Piano (Moonlight Sonata)	CD	1,2,3
Vega	Vocals (Suzanne Vega)	CD	1,2,3
Glock	Glockenspiel	EBU	1,2,3
Bassoon	Arpeggio / Melodious Phrase	EBU	2,3
Harpsichord	Arpeggio / Melodious Phrase	EBU	2,3
Soprano	Soprano singer	EBU	2,3
Guitar	Larry Coryell	EBU	2,3
Ravel	Tzigane	EBU	2,3
Strauss	R. Strauss (Orchestra)	EBU	2,3

treatments were tested, transcoded with a variety of codecs and bit rates. The treatments are listed in Table 2.

Ten expert listeners participated in the tests, which were conducted with one month intervals between experiments. The first experiment was carried out using a laptop in a sound-proofed recording studio. Listeners completed the test individually with a laptop and Superlux HD-668B, over the ear, semi-open back headphones which would be considered to be "consumer quality" (hereafter referred to as consumer). The laptop produced fan noise, raising the ambient noise in the studio from 30 to 39dBA. However, the fan noise was not reported as an issue by participants, possibly due to the partial noise shielding from the headphones. The second test repeated the MUSHRA test using the same 10 listeners listening to the same treatments but with some samples changed (as per Table 1) and using different hardware. The laptop was replaced with a Nexus 7 (2013 edition) tablet and the headphones were replaced with Sennheiser HD558, high-

Table 2: Treatments

Type	Bandwidth	Bit rate (kb/s)
reference	22 kHz/Raw-PCM	1536
anchor 1	3.5 kHz Narrowband	256
anchor 2	7 kHz Wideband	512
mp3	16kHz (SWB)	96 (CBR)
aac-he	20kHz (Fullband)	24
aac-he	20kHz (Fullband)	48
aac-he	20kHz (Fullband)	64
aac-lc	20kHz (Fullband)	128
opus	20kHz (Fullband)	128
aac-lc	20kHz (Fullband)	265

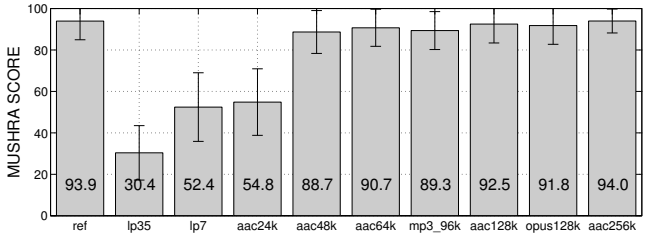

Figure 3: Consumer headphones results by treatment (for 12 music samples and 10 listeners)

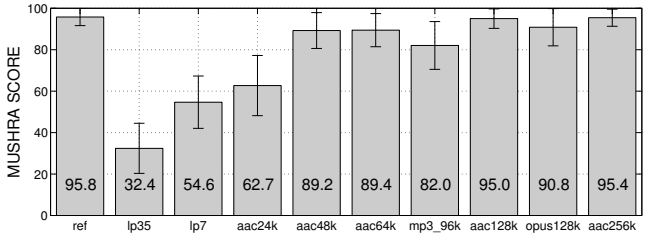

Figure 4: Studio headphones results by treatment (for 12 music samples and 10 listeners)

quality open-backed headphones which would be considered "studio quality" (hereafter referred to as studio). The use of the Nexus tablet eliminated the issue of fan noise which could have posed a problem with open-backed headphones.

A third experiment was conducted using the same materials and listeners as used in the second, but presenting the samples over loudspeakers in a recording studio (hereafter referred to as loudspeaker). The test was conducted using the Nexus 7 tablet MUSHRA test UI connected to a Mackie ONYX 1620 FireWire mixer and presented over Genelec 1031A loudspeakers. The listening point was 1.2m from the loudspeakers and the angle between the loudspeakers was 60 degrees (-30°left; +30°right). Room background noise was measured at 30 dBA and reverberation time was 0.18s at 1kHz, spatially averaged.

4. RESULTS AND DISCUSSION

Figure 2 shows the results for the six music samples that were common to all three presentation modes. The bar charts plot the treatments on the x-axis against the mean MUSHRA scores on the y-axis with 95% confidence intervals.

At first inspection, there is a trend for a modest quality improvement between presentation modes for most treatments with consumer quality headphones having the lowest quality for a given treatment and the loudspeakers having the best quality. There is also a good consistency between experiments for most treatments.

A statistical analysis was carried out using the same technique applied in audio listener tests carried out by the HydrogrenAudio forum [4], i.e. a pairwise resampling-based free step-down analysis using the max(T) algorithm [17, 11]. The consumer headphones showed no statistically significant difference between treatments for aac-he at 48 kb/s and any of the higher bit rate encodings ($p>0.099$ in all cases). For studio headphones and loudspeakers, the aac-lc 128 and 256 kb/s treatments, were still not separable from each other ($p=0.638$ and $p=0.484$ respectively) but have widened the gap over all the other treatments ($p<=0.002$ and $p<=0.033$ respectively). These results reinforce the expectation that the presentation mode is important and lower quality hardware can mask otherwise perceivable differences between treatment audio quality. However, it should be stressed that the test only compares 3 specific presentation modes labelled with indicative categories (consumer, studio, loudspeaker) and that the quality and frequency response will vary across hardware.

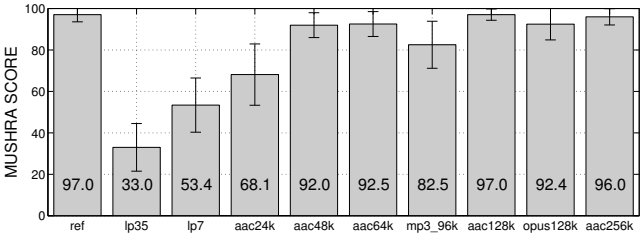

Figure 5: Loudspeaker results by treatment (for 12 music samples and 10 listeners)

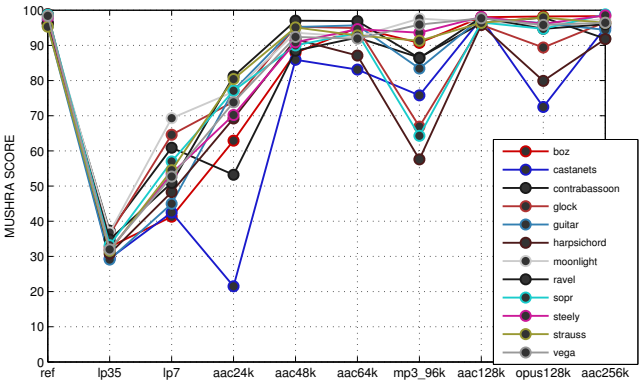

Figure 6: Loudspeaker results breakdown by music sample

Figure 3 shows result for the consumer headphones, where treatments at bit rates of 48 kb/s and above were not differentiated. Thus this experiment can be excluded for further comparison of treatment quality. Looking at the two other experiments, the results for the studio headphones follow very similar trends to those for the loudspeaker. This can be seen by comparing Figures 4 and 5. Consequently, only the results of the loudspeaker experiment will be used for a detailed analysis of treatments.

The results for the loudspeakers tested with 12 music samples are presented in Figure 5. The treatments can be broken down into a number of sub-groups. Scores for aac-he 24 kb/s were only marginally better than the 7 kHz low-pass anchor, which is not particularly surprising, given the low bit rate was not intended for music presentation. However, a better separation was evident for the higher bit rate codecs. With loudspeakers, listeners' quality perception between aac-he

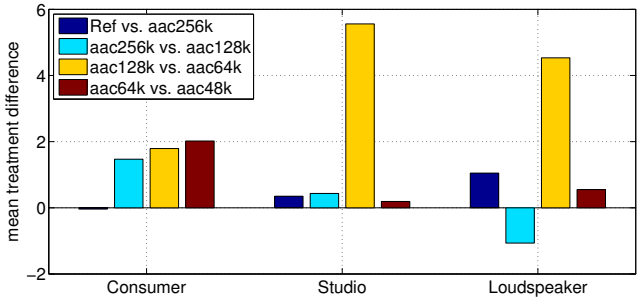

Figure 7: Average difference in MUSHRA scores between treatments per listener and sample

64 kb/s and aac-lc 128 kb/s is evident ($p<0.001$). There was no perceived quality difference for aac-he 64 kb/s over aac-he 48 kb/s ($p=0.965$) and likewise comparing the scores for aac-lc 128 kb/s to aac-lc 256 kb/s ($p=0.484$).

Figure 6 shows a breakdown of the results by sample. The trends remain largely consistent across samples. However, castanets can be seen to be an outlier for acc-he at 24 kb/s as well as MP3 and Opus. A small number of other samples exhibit the same drop in quality for MP3 and Opus as experienced with the castanets sample. This sample is a particularly difficult one for compression algorithms to deal with due to the sharp onset and rolloff times associated with the sound bursts. However, a comparison with the results from [7] showed that Opus performance for the glockenspiel was lower relative to the other samples but not for castanets. An analysis of the parameters used for the Opus transcoding highlighted that the discrepancy was due to a 5ms frame size used by Hoene et al.[7] compared to the ffmpeg default of 20ms used in these tests. Other differences in transcoding parameters were a compression level set to 0 for these tests and a constrained variable bit rate (CVBR) that was not used here. However, this change had a discernible difference in the castanets sample quality when a cursory check was carried out by the authors. The Opus developers have reported the opposite for Opus at 64 kb/s where the quality with 5 ms frames was lower than 20 ms. [15]. This highlights the importance of the codec parameters for some sample types. Conversely, many music samples are robust to a variety of codec parameters. This does call into question the magnitude of the quality difference between aac-lc 128/256 kb/s and Opus 128 kb/s as Opus can potentially be tuned to perform better with other parameters.

To assess the potential of bias due to stimulus spacing and range equalising effects [18], the results were re-analysed to investigate the mean differences between MUSHRA scores for a given treatment. This should show if there is a consistent trend in ranking one treatment higher than the other even if range equalising has occurred for individual listeners. It should also address the problem of tightly bunched stimulus spacing, i.e. having many of the treatments being much closer in quality to the reference than to the anchors. Figure 7 shows the difference between paired treatments for each experiment. The differences are small for all treatments with consumer headphones in experiment 1 and the only trend that changed in experiments 2 and 3 was the difference between aac-he 64 kb/s and aac-lc 128 kb/s.

Initial investigations show that objective measurements need to be treated with caution for low bit rate codec evaluation. Using PEAQ [12] (advanced version, Opticom gmbh), aac-he 24 and 48 kb/s are ranked equally (-3.4 ODG) but aac-lc 128 kb/s is ranked as significantly lower quality than aac-lc 256 kb/s (-2.9 and -2.2 ODG respectively).

5. CONCLUSIONS

The results for consumer headphones showed that the presentation mode can be the dominant factor over treatment type. With low quality headphones there is little difference between the codecs tested above 48 kb/s. The results were bimodal for some codecs and sample types, highlighting that the type of music can be an important factor in whether there is a perceptible quality drop for a given treatment. The parameters used in configuring the codec also matter (i.e. Opus frame size), although the conflicting results point to further investigation being required here. These results indicate that significant bandwidth savings can be achieved given knowledge of listening equipment at the decoder. Practically, if headphones provided a hardware ID-tag input to the playback device and measured the ambient background noise level, the streaming audio codec and bit rate could be adjusted according to the environment. In summary, the choice of codecs and bit rate do matter, but only if the listening equipment is above a quality threshold. These results provide useful insights into the influence of presentation mode, audio sample content, specific codec parameters and bit rates which will inform ongoing work on objective audio metric development, specifically the adaptation of ViSQOL [5, 6] for audio quality estimation.

6. REFERENCES

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

[2] European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). Qualinet White Paper on Definitions of Quality of Experience, 2012.

[3] B. Feiten, J. Kroll, A. Raake, M. Wältermann, and U. Wüstenhagen. Evaluation of super-wideband speech and audio codecs. In *Audio Engineering Society Convention 129*, Nov 2010.

[4] H. A. W. Forum. http://hydrogenaudio.org.

[5] A. Hines, J. Skoglund, A. Kokaram, and N. Harte. VISQOL: The Virtual Speech Quality Objective Listener. In *IWAENC*, 2012.

[6] A. Hines, J. Skoglund, A. Kokaram, and N. Harte. Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[7] C. Hoene, J.-M. Valin, K. Vos, and J. Skoglund. Summary of opus listening test results – internet-draft. *http://tools.ietf.org/html/draft-ietf-codec-results*, 2012.

[8] ISO/IEC 13818-7:2006. Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC), 2006.

[9] G. Lorho. *Perceived quality evaluation: an application to sound reproduction over headphones*. PhD Thesis, Aalto University, Finland, 2010.

[10] D. Marston and A. Mason. Cascaded audio coding. *EBU Technical Review*, 2005.

[11] G.-C. Pascutto. Bootstrap. www.sjeng.org/bootstrap.html, 2011.

[12] Rec. ITU-R. BS.1387. Perceptual Evaluation of Audio Quality (PEAQ). *International Telecommunication Union*, 1998.

[13] Rec. ITU-R. BS. 1534-1. Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA). *International Telecommunication Union*, 2003.

[14] D. Schobben and S. van de Par. The effect of room acoustics on mp3 audio quality evaluation. In *Audio Engineering Society Convention 117*, Oct 2004.

[15] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos. High-quality, low-delay music coding in the opus codec. In *Audio Engineering Society Convention 135*, 2013.

[16] J.-M. Valin, K. Vos, and T. Terriberry. Definition of the Opus Audio Codec. *IETF*, September, 2012.

[17] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley Series in Probability and Statistics*. John Wiley and Sons, New York, January 1993.

[18] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey. Potential biases in mushra listening tests. In *Audio Engineering Society Convention 123*, 2007.