Audio Engineering Society

Convention Paper 10001

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

A subjective evaluation of high bitrate coding of music

Kristine Grivcova[1], Chris Pike[1], and Thomas Nixon[1]

[1]*BBC Research & Development*

Correspondence should be addressed to Kristine Grivcova (kristine.grivcova@bbc.co.uk)

ABSTRACT

The demand to deliver high quality audio has led broadcasters to consider lossless delivery. However the difference in quality over formats used in existing services is not clear. A subjective listening test was carried out to assess the perceived difference in quality between AAC-LC at 320kbps and an uncompressed reference, using the method of ITU-R BS.1116. Twelve audio samples were used in the test, which included orchestral, jazz, vocal music, and speech. A total of 18 participants with critical listening experience took part in the experiment. The results showed no perceptible difference between AAC-LC at 320 kbps and the reference.

1 Introduction

Popular demand for high quality content and increasing bandwidth for audience delivery have sparked broadcasters' interest in delivery using lossless codecs, which exactly preserve the original audio signal. The Free Lossless Audio Codec (FLAC) is a widely used lossless codec. Some streaming services such as Tidal are already offering lossless streaming [1] and other companies like Spotify are publicly investigating the prospect [2].

BBC Radio 3 is a radio station with a focus on classical music and opera; it also features other content such as jazz, world music, drama, culture and the arts. It is a service that distinguishes itself by its sound quality and delivers content in stereo at high bitrates, up to 320 kbps AAC-LC (Advanced Audio Codec, Low Complexity profile). BBC Radio 3 has been investigating provision of a lossless streaming option to listeners. With this, the question has been raised as to whether the increase in quality is perceptible and beneficial to the listeners.

There are few studies evaluating the quality of AAC coding of stereo signals at such high bitrates. Previous work has mainly focused on lower bitrates (up to 128 kbps) [3, 4] or coding of multichannel signals [5, 6]. Those studies that evaluated coding of stereo signals with AAC at 128 kbps followed the test method of Recommendation ITU-R BS.1116 [7] and showed that participants found the quality difference perceptible but not annoying [3, 4].

Coding of stereo signals with a sampling rate of 96 kHz and bit depth of 24 bit was investigated in [8], also using the ITU-R BS.1116 evaluation method. Here perceptual transparency was observed for the AAC coded version using bitrates in the range 160-256 kbps. In [9], the quality of MP3 at 256 kbps was

compared to WAV (44.1 kHz, 16 bit), which showed that there was no perceptible difference in quality. However the test did not follow a standardised method, indirect assessment of differences was made using audio production tasks, which may have been less sensitive.

A formal listening test was conducted to assess the perceived difference in quality between stereo signals coded with AAC-LC at 320 kbps and an uncompressed reference. The test was run in the BBC R&D listening room which complies with Recommendation ITU-R BS.1116 [7]. The test design and the obtained results are presented in this paper.

2 Codecs under investigation

This section describes the codecs involved in this study in more detail.

2.1 Free Lossless Audio Codec

FLAC is a lossless codec, meaning it preserves the original signal and it will not introduce artefacts. Bitrates of FLAC encoded signals vary depending on the content, but the codec generally reduces the data rate by 50–60% [10]. FLAC allows to store the original metadata from WAV files and add additional metadata. Since the decoded signals would be identical, the reference stimuli in the listening test were the uncompressed original versions.

2.2 Advanced Audio Codec

AAC is a lossy codec, which means that the audio signal is modified. There may be a quality loss, with audible artefacts introduced. The likelihood of audible artefacts is dependent upon the bitrate, at higher bitrates audible artefacts will become unlikely, with lower bitrates they are expected. The likelihood of artefacts is also dependent upon the format of the input signals and the nature of the signal content. Common artefacts when using the Low Complexity (LC) profile include loss of high frequency content, pre-echo, distortion and aliasing [11]. Quality loss also appears in stereo imaging which gets increasingly worse with lower bitrates where joint-stereo or spatial audio coding is often used.

The High Efficiency (HE) profile uses spectral band replication (SBR) to improve the efficiency of representing high frequency content; however it can also add new artefacts such as tone trembling, tone shift, noise overflow and beat effect [12].

For this test the Fraunhofer FDK AAC library was used. This is the same encoder used for internet distribution of BBC Radio 3. A range of bitrates from 48 to 320 kbps were used with two different AAC profiles (LC and HE) to encode the stimuli included in the test and training phase, as well the item selection process, which will be explained in more detail in Section 3.

3 Test material selection and preparation

Some audio content is more challenging for the codecs than other. For example, a recording of percussive sounds may be more prone to revealing pre-echo artefacts because it affects mainly transients. Instruments with complex spectral content such as the violin are prone to tone trembling and tone shift artefacts. Orchestral recordings are more prone to revealing stereo image alteration, in addition to other artefacts. Such knowledge was used to inform the selection process for the audio items to use in the test.

3.1 Initial material selection

Initially 34 items were selected from various pre-recorded BBC Radio 3 programmes. The programmes were selected to be representative of the station output and likely to be challenging for the codecs.

Since the radio programmes are typically at least a half an hour long, shorter clips were extracted. The duration of the clips was 10–25s as specified in Recommendation ITU-R BS.1116 [7]. Start and end points were set appropriately so as not to distract the test participant.

3.2 Final test material selection

The initial selection of 34 samples was reduced to 12 samples [7] in order to fit into the specified time frame (20–30 minutes per session) and avoid listener fatigue. This was achieved through a listening session, where

the experimenters reviewed all 34 items with multiple different encodings. Each item was encoded using the AAC-LC profile at bitrates of 64 kbps, 128 kbps, and 320 kbps. A range of bitrates was used because it was understood that artefacts may not be audible at 320 kbps. This process aimed to reveal the most sensitive items, where artefacts were clearly audible at lower bitrates. After coding, the samples were time aligned and loudness aligned to -23 LUFS [13]. This process was used for the item selection process and for the listening test itself. The codec settings used for the listening test are described in Section 4.3.

During the selection process, all items were assigned a sensitivity rating between one and three, with the following definitions:

- 1 – very obvious impairments at 64 and 128 kbps, potentially audible artefacts at 320 kbps;
- 2 – obvious impairments at 64 kbps, barely audible artefacts at 128 kbps, nothing noticeable at 320 kbps;
- 3 – impairments hard to detect at 64 kbps, no audible artefacts at 128 and 320 kbps.

Twelve items were assigned a rating of one and were included in the final test. The selected items are listed in Table 1.

Item name	Description
electric_guitar	Guitar, clarinet, electric guitar
forest_chant	Vocals with backing music, songs from forests in Cameroon, strings,
harpsichord	Recorder, harpsichord
instrumental_percussion	Orchestra
jazz_percussion	Saxaphone, piano, bass
live_speech_male	Male voice, applause
orchestra_2	Symphony orchestra
orchestra_3	Symphony orchestra
percussion_2	Percussion, marimba, vibraphone
piano_2	Piano
piano_strings	Piano quintet, string quartet
strings_2	Clarinet and string quartet

Table 1: The final selection of test items

4 Test design

The test design followed Recommendation ITU-R BS.1116-3 [7].

4.1 Test method

The test used the double-blind triple stimulus with hidden reference presentation method. On each test page, the listener was presented with three stimuli: the reference, and two test stimuli labelled A and B. The reference was the uncompressed version of the item and A and B were randomly assigned to either a hidden reference version or the processed version of the item. Each of the 12 test items was presented to the participants on separate pages, they were asked to rate the stimuli using the ITU five-grade impairment scale shown in Table 2. The items were presented in a random order to the participants.

The listeners could freely switch between stimuli and listen to each item for as long as they wished. A slider was available for stimuli A and B to assign the ratings.

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Table 2: ITU five-grade impairment scale used for the test [7]

An initial pilot test was run with the 12 selected items, which were encoded with AAC-LC at 320 kbps. It showed that assessors often had difficulty detecting any differences. To be able to assess discrimination ability of the assessors it was decided to also present the 12 items with processed stimuli coded using HE-AAC at 48 kbps. It was found that HE-AAC 48 kbps provided

reasonably high quality output where, to the untrained listener, the impairments would not be immediately obvious but experienced listeners could reliably determine differences. Hence it was decided to use it to check the reliability and consistency of the listeners.

4.2 Test structure

The participants were presented with written instructions explaining the test method and structure. They first performed a familiarisation exercise, which involved listening to all of the audio test material and learning how to use the rating interface. After that they carried out the grading process, where the results were recorded for later analysis.

The grading phase involved 24 rating trials (12 items and 2 codec settings). The test was split into two sessions. The first session involved the training phase and the first 12 grading trials. The second session involved the final 12 grading trials. The sequence and codec settings of the items was randomised for each participant.

To allow sufficient time for each session and account for different pace of each participant, only four sessions were scheduled per day. The participants were encouraged to choose both sessions on the same day; however, this was not always possible due to busy work schedules. For the participants doing their second session on a different day, additional training items were added to the second session. This allowed the listeners to refresh their memory of the task ahead and tune their ears to listening for very small impairments.

In the instructions provided to the participants, they were encouraged to avoid guessing and leave both sliders at the maximum rating of five if they could not perceive the difference in quality.

4.3 Training phase

The training phase is an important part of the test which allows the participants to become familiar with the content and potential artefacts they will be listening for, as well as adjusting to the listening conditions and learning to use the test software.

In this case the training phase consisted of the 12 items also used in the grading phase. The listener was presented with five stimuli of which one was the declared uncompressed reference and those labelled A, B, C and D of which three were processed (AAC encoded) versions of the item and one was a hidden reference, all assigned randomly for each participant. The training items were presented in the same order to all the listeners.

The three processed versions were encoded at AAC-LC 48 kbps, HE-AAC 48 kbps and AAC-LC 320 kbps. The HE-AAC 48 kbps and AAC-LC 320 kbps settings were the same as in the grading phase of the test. Additionally, AAC-LC at 48 kbps was used as it gave low quality. This ensure some stimuli had more obvious artefacts during the training process.

The participants were encouraged to attempt to rate the items the same way they would in the grading phase.

4.4 Listening panel

A total of 18 participants with experience in critical listening took part in the listening test. Participants had various backgrounds including broadcast sound engineers, R&D engineers and radio operations engineers. Thirteen participants had significant experience in listening tests and the rest had no prior listening test experience but had experience in other types of critical listening tasks, such as audio mixing. The pool of participants consisted of 4 females and 14 male listeners.

5 Results

In this section, the results of the test are presented. Prior to the analysis the following questions were set.

- Are the listeners reliable enough and not giving random answers?
- Did the two-session approach affect the results?
- Is AAC-LC 320 kbps encoded material distinguishable from the lossless versions?
- How does program material affect the codec performance?
- Are there any other unexpected results to report?

The aim was to answer these questions using statistical analysis methods applied to the obtained results. Throughout the analysis process the difference grades of the results were used, shown in Table 3. The difference grades were calculated by subtracting the grade of the hidden reference from the grade of the coded stimulus. This allowed using a single number to reflect whether the participant had marked down the coded version or the reference. If the difference grade is negative, it means the coded version was downgraded and if it is positive, the reference has been downgraded.

Impairment	Grade	Diffgrade
Imperceptible	5.0	0.0
Perceptible, but not annoying	4.0	-1.0
Slightly annoying	3.0	-2.0
Annoying	2.0	-3.0
Very annoying	1.0	-4.0

Table 3: Difference grades when the coded stimulus is downgraded on the ITU five-grade impairment scale

5.1 Post-screening

To assess whether the listeners have given reliable data, ITU-R BS.1116 recommends a post-screening process. This allows to determine whether each listener can really hear the impairments or is merely guessing. The results of AAC-LC 320 kbps encoded material were not included in post-screening analysis as the audio quality is such that it is difficult to determine if the artefacts are present and therefore would not reflect the critical listening ability of the participants.

Instead the results from the HE-AAC 48 kbps material were analysed to check that the assessors could reliably differentiate it from the reference. The tests aimed to determine if the scores given by the participants were consistently below zero to indicate their ability to hear artefacts. In this case data was not normally distributed and therefore t-test can be unreliable, so a Wilcoxon test was used in addition to the t-test to validate the results. The alternative hypothesis was set to $\mu < 0$ with a significance level of 5%. The p-values for each participant from both tests are presented in Table 4, similar results are obtained from both tests.

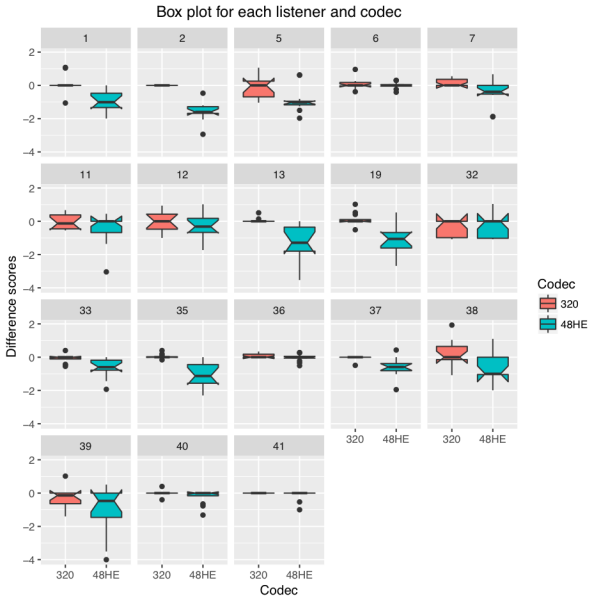

Fig. 1: Box plots of the difference scores between the hidden reference and the codecs under test, for each participant. Notches show bootstrapped 95% confidence intervals of the median.

In addition, Table 4 shows how many times the hidden reference was downgraded by each participant (error count and error percentage). The allowed error rate in this case was 25% or 3 errors. The distribution of results from each listener was also inspected using box plots, as shown in Figure 1.

Five participants were removed during the post-screening process, indicated by the highlighted rows in Table 4. Participants 6, 12, and 36 were removed due to t-test results and the number of errors. In addition participants 40 and 41 were remove as their scores suggested inability to detect artefacts in the stimuli coded with HE-AAC at 48kbps.

5.2 Overall results

After the post-screening process the scores of the remaining 13 participants were used for further analysis. This section will attempt to answer the question asked prior to the listening test: can the difference in quality between lossless and AAC-LC 320 kbps encoding be perceived? Table 5 presents

ID	t-test	Wilcoxon	Errors	Error %
1	0.0002	0.0046	0	0
2	0.0000	0.0019	0	0
5	0.00105	0.0027	2	17
6	0.473	0.0542	3	25
7	0.0529	0.118	2	17
11	0.0629	0.0541	2	17
12	0.193	0.305	3	25
13	0.000907	0.00451	0	0
19	0.000736	0.00333	1	8
32	0.0956	0.1704	1	8
33	0.00151	0.00292	0	0
35	0.000241	0.00296	0	0
36	0.346	0.338	3	25
37	0.00274	0.00402	1	8
38	0.0245	0.0413	2	17
39	0.0230	0.0261	2	17
40	0.0499	0.0907	0	0
41	0.0937	0.186	0	0

Table 4: Post-screening results of each participant for the HE-AAC 48kbps scores including p-values from t-test and Wilcoxon test, as well as error count. Highlighted rows indicate assessors removed during post-screening.

the results for AAC-LC 320 kbps encoding for all listeners in terms of a mean and a t-test p-value. A one sided one sample t-test was used with significance level of 5%. The null hypothesis was set to μ=0 and alternative hypothesis was set to μ <0. For the null hypothesis to be accepted the mean of the scores would have to be around zero, which would suggest no audible difference between AAC-LC 320 kbps and uncompressed signal. The alternative hypothesis would be accepted if the mean of the scores were significantly different, in this case less than zero, which would suggest that there was an audible difference between the codecs.

Table 5 shows the mean of AAC-LC 320 kbps encoding is very close to zero. The p-value from the t-test shows that the scores are not statistically different from zero. This indicates that participants were not able to perceive the difference between lossless and AAC-LC 320 kbps encoding.

Box plots of the distribution of difference grades from

Codec	Mean	p-value	t-value
AAC-LC 320 kbps	-0.0145	0.351	-0.384
AAC-HE 48 kbps	-0.786	1.0214e-21	-10.98

Table 5: t-test result for all scores for both codecs

all listeners for both HE-AAC 48 kbps and AAC-LC 320 kbps are shown in Figure 2. The scores of the HE-AAC 48 kbps items have much larger range and are mostly below zero, which reinforces the result of the t-test (p << 0.05). The median result for the HE-AAC 48 kbps falls in range from 0 to -1, which relates to 'perceptible but not annoying' on the grading scale.

The AAC-LC 320 kbps scores show a number of outliers. Since it was difficult to hear differences, some listeners were more prone to making mistakes in identifying artefacts. However, the errors mostly fall into 'perceptible but not annoying' category, which means even if the listeners thought they could hear an artefact it was not substantially affecting their experience. The outliers appear both above and below zero, which supports the suggestion that they were guessing.

5.3 Other results

An analysis of variance (ANOVA) was used to assess the effect of sessions, codecs and items on the results. The chosen significance level was 5%.

The results show that the session did not have a statistically significant effect on the scores. This means results from both sessions could be combined and analysed together.

The results show that the items had a statistically significant effect on the scores, confirming that different material is affected by the codecs differently. This prompted a further investigation into which items are more sensitive and are affected more by the codecs.

Figure 3 shows the results for each item using box plots. The first item that is worth mentioning is *live_speech_male*, which has significantly lower scores at HE-AAC 48 kbps than other items. This is because it contained applause along with male speech, which was challenging for the codec and artefacts were

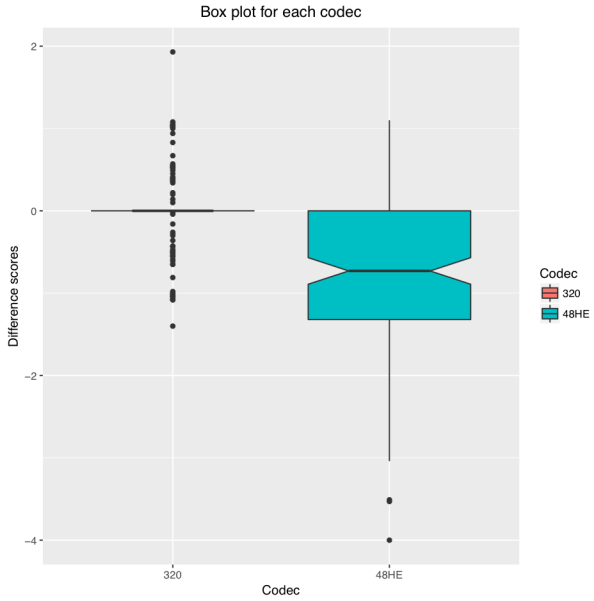

Fig. 2: Box plots of the difference scores between the hidden reference and the codecs under test, for each codec, across all listeners. Notches show bootstrapped 95% confidence intervals of the median.

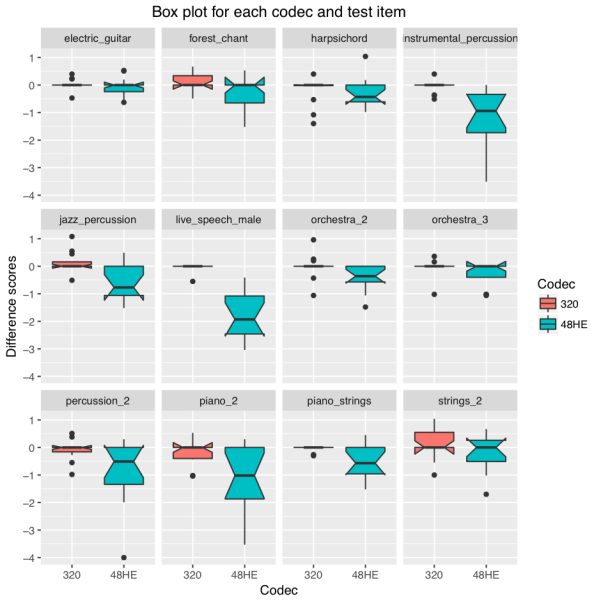

Fig. 3: Box plots of the difference scores between the hidden reference and the codecs under test, for each programme item

obvious to the listeners.

Another interesting item to consider is *piano_2*, for which the median is zero, but there is a slight negative skew. This suggests that some listeners might have perceived the difference, but did not think it was annoying. However it would require additional testing to obtain a clearer result. This item was also challenging for the HE-AAC 48 kbps encoding, where it received scores indicating 'slightly annoying'.

The plot of the results for the item *strings_2*, showed the largest number of errors with a slight positive skew in distribution for 320kbps AAC-LC. This was potentially due the original recording being noisy, which might have confused the listeners into thinking they could hear coding artefacts.

6 Summary

This paper presented a subjective listening test to determine whether there is likely to be a perceptible difference between lossless (FLAC) and AAC 320 kbps coding. Recommendation ITU-R BS.1116-3 was used as guideline for the design process. A total of 18 participants took part in the test and each graded 12 test items on the ITU five-grade impairment scale. The results were analysed using difference grades with statistical methods, such as the *t*-test and ANOVA. The post-screening process was used to eliminate the scores of five participants.

The analysis showed that there was no statistically significant difference in quality between the uncompressed signals and AAC-LC 320 kbps coding, which means participants generally could not perceive differences between the two versions. It also showed that there was a statistically significant difference between the uncompressed signals and HE-AAC 48 kbps coding. This means participants could perceive differences in quality between these two versions.

The results show that AAC coding can preserve the quality of the original audio at high bitrates. This suggests that offering lossless audio might not have a great benefit in terms of quality increase to the consumers. However to ensure that a delivery service is transparent, with original quality always maintained

for all signals, a lossless codec may still be required.

References

[1] Tidal, "About Tidal," *About TIDAL*, 2018, http://tidal.com/lp/about/.

[2] Bi, F. and Marino, A., "Spotify is testing lossless audio. Can you hear the difference?" *The Verge*, 2017, https://www.theverge.com/2017/4/5/15168340/lossless-audio-music-compression-test-spotify-hi-fi-tidal.

[3] Ehret, A., Kjörling, K., Rödèn, J., Purnhagen, H., and Hörich, H., "aacPlus, only a low-bitrate codec?" in *AES 117th Convention in San Francisco, CA, USA*, Audio Engineering Society, 2004.

[4] Soulodre, G. A., Grusec, T., Lavoie, M., and Thibault, L., "Subjective Evaluation of State-of-the-Art 2-Channel Audio Codecs," in *AES 104th Convention in Amsterdam*, Audio Engineering Society, 1998.

[5] EBU Tech 3324, "EBU Evaluations of Multichannel Audio Codecs," Technical report, EBU, 2007.

[6] Sugimoto, T., Nakayama, Y., and Oode, S., "Bit Rate of 22.2 Multichannel Sound Signal Meeting Broadcast Quality," in *AES 137th Convention in Los Angeles, CA, USA*, Audio Engineering Society, 2014.

[7] ITU Radiocommunication Sector, "Method for the subjective assessment of small impairments in audio systems," *Recommendation ITU-R BS.1116-3*, 2015.

[8] Bürgel, C., Bartholomäus, R., Fiesel, W., Hilpert, J., Hölzer, A., Linzmeier, K., and Weishart, M., "Beyond CD-Quality: Advanced Audio Coding (AAC) for High Resolution Audio with 24 bit Resolution and 96 kHz Sampling Frequency," in *AES 111th Convention*, New York, 2001.

[9] Martin, D., King, R., Woszczyk, W., Massenburg, G., and DeFrancisco, M., "Can We Hear the Difference? Testing the Audibility of Artifacts in High Bit Rate MP3 Audio," in *AES 141st Convention in Amsterdam*, Audio Engineering Society, 2016.

[10] Coalson, J., 2018, https://xiph.org/flac/comparison.html.

[11] Dick, S., Schinkel-Bielefeld, N., and Disch, S., "Generation and Evaluation of Isolated Audio Coding Artifacts," in *AES 143rd Convetion in New York, NY, USA*, Audio Engineering Society, 2017.

[12] Liu, C.-M., Hsu, H.-W., Yang, C.-H., Tang, S.-H., Lee, K.-C., Yang, Y.-C., Chang, C.-M., and Lee, W.-C., "Compression Artifacts in Perceptual Audio Coding," in *AES 121st Convetion in San Francisco, CA, USA*, Audio Engineering Society, 2006.

[13] ITU Radiocommunication Sector, "Algorithms to measure audio programme loudness and true-peak audio level," *Recommendation ITU-R BS.1770-4*, 2015.