Research & Development
White Paper

WHP 384

May 2018

A subjective evaluation of high bitrate coding of music

K Grivcova, C Pike, T Nixon

Design & Engineering

BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 384

A subjective evaluation of high bitrate coding of music

Kristine Grivcova Chris Pike Tom Nixon

Abstract

The demand to deliver the highest quality audio has pushed broadcasters to consider lossless delivery. However, there is a lack of existing perceptual test results for the codecs at high bitrate. Therefore, a subjective listening test (ITU-R BS.1116-3) was carried out to assess the perceived difference in quality between AAC-LC 320kbps and an uncompressed reference. Twelve audio samples were used in the test, which included orchestral, jazz, vocal music and speech. A total of 18 participants with various experience levels took part in the experiment. The results showed no perceptible difference between lossless and AAC-LC 320 kbps encoding.

This paper was originally presented at the 144th Convention of the Audio Engineering Society, 23–26 May 2018 in Milan, Italy and is also available from the AES's electronic library at URL: http://www.aes.org/e-lib/browse.cfm?elib=19397.

Additional key words:

BBC Research & Development
White Paper WHP 384

A subjective evaluation of high bitrate coding of music

Kristine Grivcova Chris Pike Tom Nixon

1 Introduction

Ever increasing demand for high quality content has sparked interest of the broadcasting world in delivery using lossless encoding, which preserves the original quality of the audio. One coded that is used in the industry is Free Lossless Audio Codec (FLAC). Some streaming services such as Tidal are already offering lossless quality streaming [1] and other companies like Spotify are publicly investigating offering it [2].

BBC Radio 3 is a station with focus on classical music and opera; but 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. Therefore, BBC Radio 3 has also been investigating provision of a lossless streaming option to their listeners. With this, the question has been raised as to the increase in quality is perceptible and beneficial to the listeners.

There is a lack of previous work in evaluating AAC at high bitrates as it has mainly been focused on lower (up to 128kbps) bitrate audio quality [3, 4] or multichannel audio quality [5]. The most relevant work was presented in [6] and compared audio quality of MP3 256kbps to WAV (44.1k, 16bit), which showed that there is no perceptible difference in quality. However, the test did not follow any standardised test method. Other work that has followed ITU-R BS.1116-3 guidelines to test audio quality at 128kbps stereo showed that participants found the quality difference perceptible but not annoying [3, 4].

Therefore, a formal listening test was conducted to assess the perceived difference in quality between stereo signals coded with AAC Low Complexity (AAC-LC) profile 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.1116-3 [7]. The test design and the obtained results are presented in this paper.

2 Codecs used in the test

In this section the Advanced Audio Codec (AAC) will be described, which was used to encode the test stimuli using various bitrates. Since FLAC is the target encoder for lossless delivery it will also be briefly described in this section.

2.1 FLAC

FLAC is an encoder that preserves the original quality of the recording, it will not introduce artefacts. It reduces the data rate by 50–60% [8]. It allows to store the original metadata from WAV files and add additional metadata. The bitrates of FLAC encoded signals can vary depending on the content. Since the FLAC codec does not modify the input signals, the reference stimuli were the uncompressed original versions.

2.2 Advanced Audio Codec

AAC is a lossy encoder which means that the audio signal is modified by the codec. There may be a quality loss, with audible artefacts introduced. The likelihood of audible artefacts is dependent upon the bitrate. Common artefacts when using Low Complexity (LC) profile include loss of high frequency content, pre-echo which appears due to spreading of the noise around transients, distortion and aliasing [9]. Quality loss also appears in stereo imaging which gets increasingly worse with lower bitrates due to joint-stereo or spatial audio coding.

High Efficiency (HE) profile uses spectral band replication (SBR) to improve the efficiency of representing high frequency bandwidth; however it can also add new artefacts such as tone trembling, tone shift, noise overflow and beat effect [10].

For this test the Fraunhofer FDK AAC codec 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 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 coders than other. For example, a recording of percussive sounds will be more prone to revealing pre-echo artefact because it affects mainly transients. Instruments with complex spectral content such as 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. This information was used to inform the selection process of the test items.

3.1 Initial material selection

Initially 34 items were selected from various BBC Radio 3 pre-recorded 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 between 10–25s as specified in Recommendation ITU-R BS.1116-3 [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 of 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 artefacts may not be audible at 320kbps. This process aimed to reveal the most sensitive items, where artefacts were clearly audible at lower bitrates. After coding, the samples were time and level aligned to -23LUFS [11]. This is the same process for pre-selection and the test. The encoding settings are described in more detail in Section 4.2.

During the selection process, all items were assigned a sensitivity rating between one and three where: 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. As a result, 12 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 follows Recommendation ITU-R BS.1116-3 [7]. The participants were presented with written instructions explaining the structure of the test. They first performed a familiarisation exercise, which involved listening to all of the audio test material. After that they carried out the grading process, where the results were recorded and later used for analysis.

In the grading phase the listeners were asked to compare basic audio quality between the reference (uncompressed version) and two stimuli, of which one was AAC encoded and the other a replica of the reference. Each of the 12 test items was presented to the participants, they were asked to rate the stimuli using the ITU five-grade impairment scale shown in Table 2. The test was run across two sessions, where the first session included the training and grading phase and the second session only included the grading phase.

4.1 Test structure

The test used the double-blind triple stimulus with hidden reference method. On each test page, the listener was presented with three stimuli: the Reference, and two test stimuli labelled A and B. In this scenario, the reference was the uncompressed version and A and B were randomly assigned either hidden reference or the processed version of the item. 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 rating.

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

Table 2: Rating scale used for the test [7]

An initial pilot test was run with the 12 selected items, which were encoded with AAC-LC 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 at HE-AAC 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 listener.

The grading phase involved 24 rating trials (12 items and 2 codec settings). The test was split into two sessions. The session first 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 12 test items. 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.

4.2 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 labeled 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 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. HE-AAC 48 kbps and AAC-LC 320 kbps bitrates were chosen to represent what could be expected in the test in terms of artefacts. Additionally, AAC-LC 48 kbps was selected as very low quality to make the artefacts obvious during the training.

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

4.3 Listening panel

A total of 18 participants with experience in critical listening took part in the listening test. Participants had various backgrounds including studio managers, R&D engineers and radio operations engineers. Thirteen participants had 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 Result analysis

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 grade is calculated as the grade of the coded stimuli minus the grade of the hidden reference. 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 meant 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: Grading scale with negative diffgrade example when coded stimuli is downgraded

5.1 Post-screening

To assess wether the listeners have given reliable data Recommendation ITU-R BS.1116-3 [7] suggests 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.

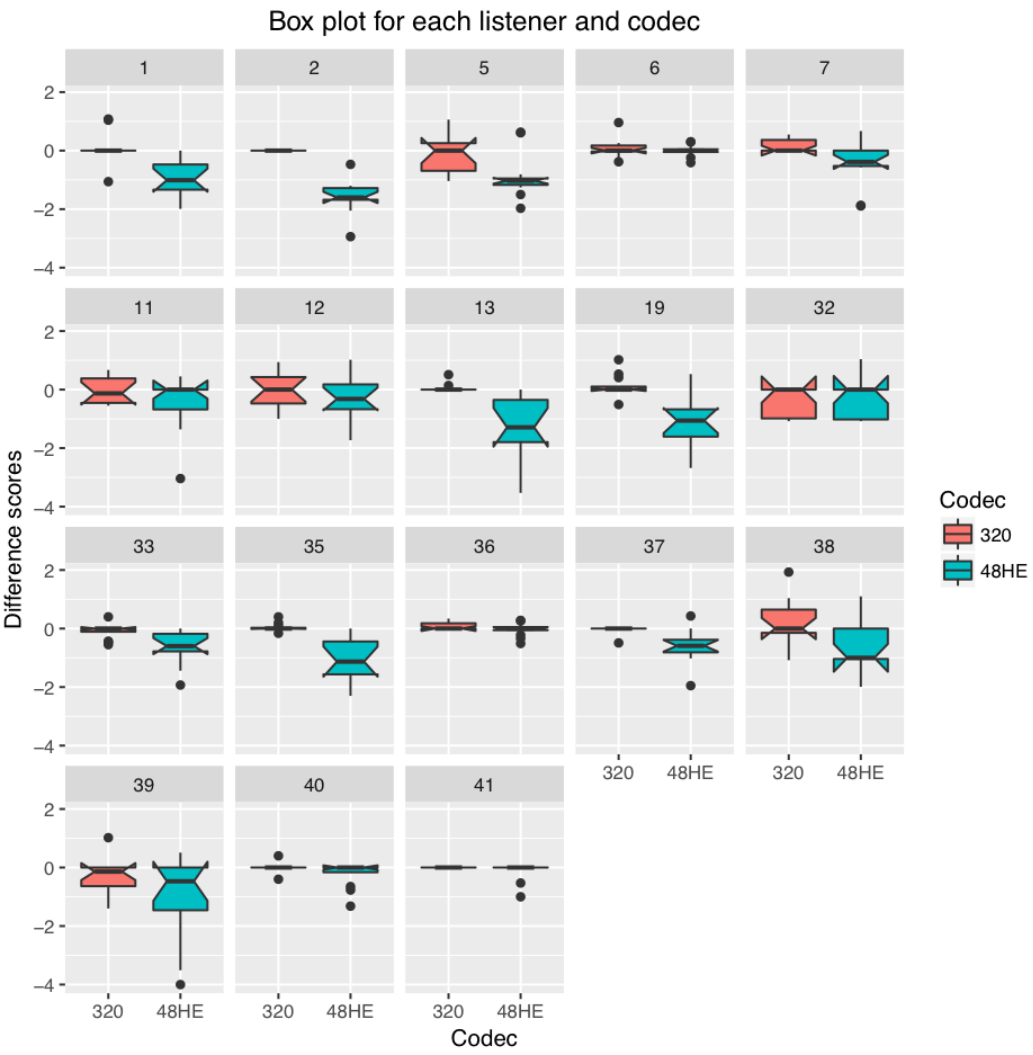

Figure 1: Box plots of the difference scores between the hidden reference and the codecs under test, for each participant

Instead the results from the lower anchor 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. 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 and confirm the validity of the t-test results as they are not notably different.

In addition, Table 4 shows how many times the hidden reference was downgraded by each participant (error count and error percentage). The highlighted rows in Table 4 indicate the participants which were removed as result of the screening process. The allowed error rate in this case was 25% or 3 errors.

Five participants were removed during the post-screening process. 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 determine artefacts in the audio encoded with

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.

HE-AAC 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 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 will be 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.

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

The results for HE-AAC 48 kbps and AAC-LC 320 kbps codec were plotted using difference grades represented by box plots 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”.

Many of AAC-LC 320 kbps codec scores are outliers, which is due to the fact that it was extremely difficult to hear difference and listeners were more prone to making mistakes. 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.

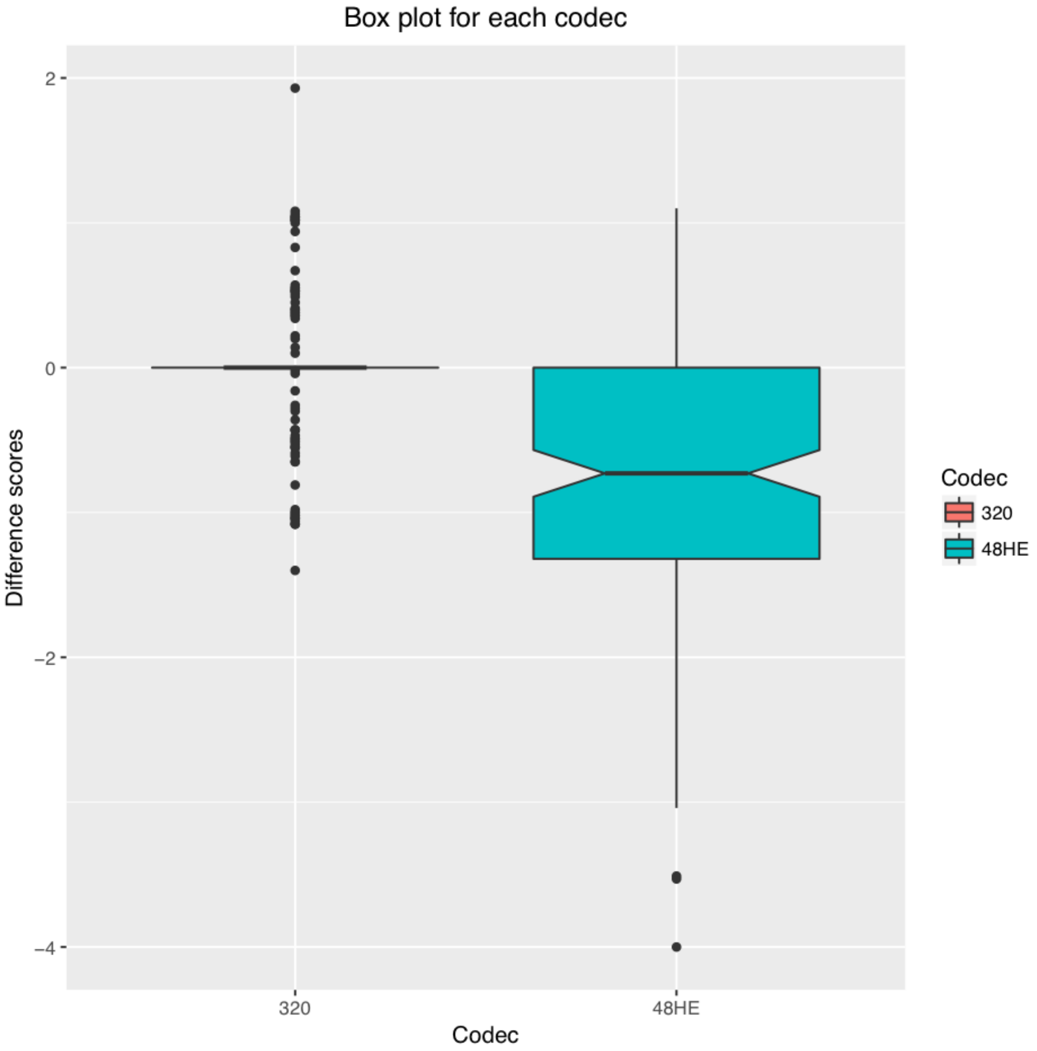

Figure 2: Box plots of the difference scores between the hidden reference and the codecs under test, for each codec

5.3 Other results

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 items have had a statistically significant effect on the scores, confirms 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 any other item. This is because it contains applause along with male speech which causes a lot of problems for the codec and therefore artefacts were very obvious to the listeners.

Another interesting item to consider is *piano_2*, of which 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 more certain result. This item was also challenging for the HE-AAC 48 kbps encoding, where it has received scores in the area classed as 'slightly annoying'.

The plot of the results of 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.

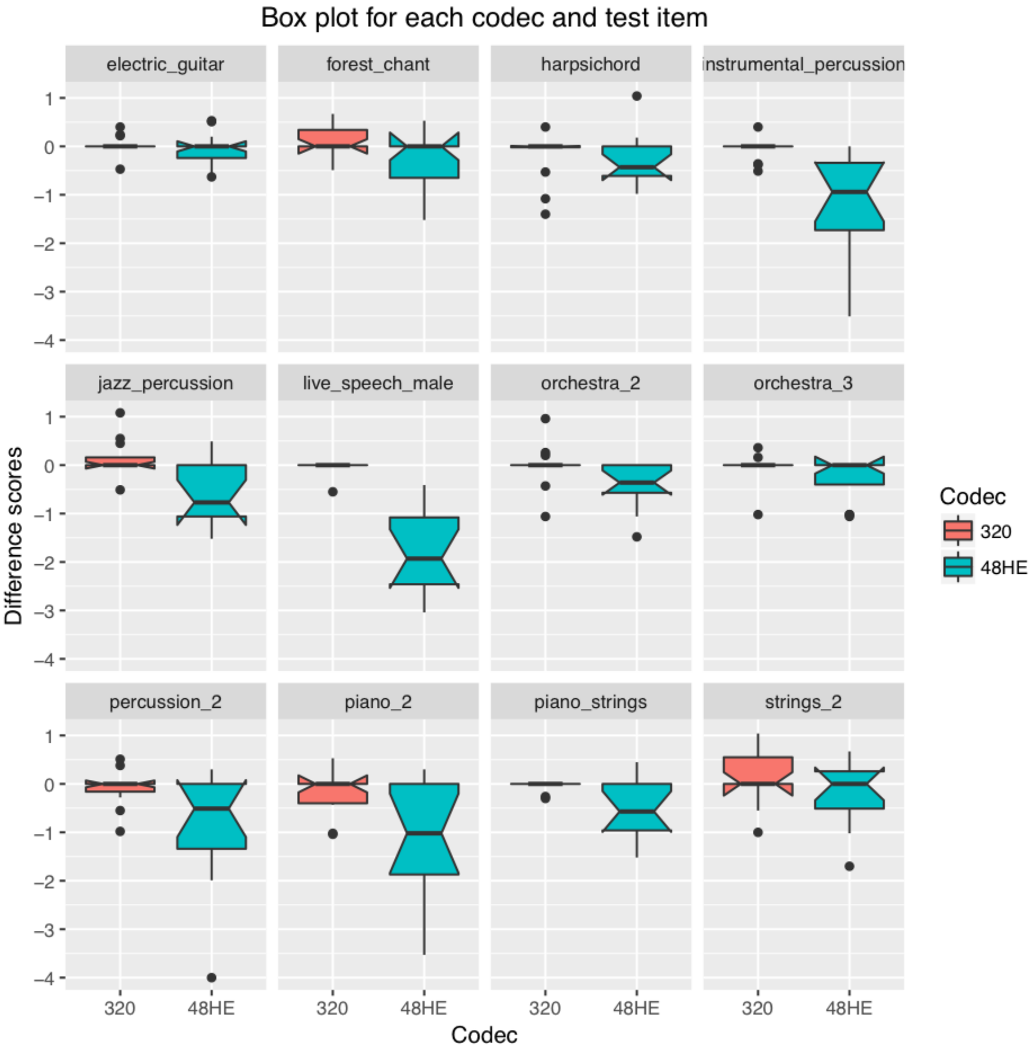

Figure 3: Box plots of the difference scores between the hidden reference and the codecs under test, for each programme item

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 compression. 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 5-grade impairment scale. The results were analysed using difference grades with statistical methods, such as *t*-test and ANOVA. The post-screening process was used to eliminate the scores of 5 participants.

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

The test has shown how AAC encoders can preserve the quality of the original audio. 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 and original quality is always maintained, a lossless codec would be required.

References

[1] Tidal. “About Tidal”. In: *About TIDAL* (2018). http://tidal.com/lp/about/.

[2] Frank Bi and Andrew Marino. “Spotify is testing lossless audio. Can you hear the difference?” In: *The Verge* (Apr. 2017). https://www.theverge.com/2017/4/5/15168340/lossless-audio-music-compression-test-spotify-hi-fi-tidal.

[3] Andreas Ehret et al. “aacPlus, only a low-bitrate codec?” In: *AES 117th Convention in San Francisco, CA, USA*. Audio Engineering Society, Oct. 2004.

[4] Gilbert A. Soulodre et al. “Subjective Evaluation of State-of-the-Art 2-Channel Audio Codecs”. In: *AES 104th Convention in Amsterdam*. Audio Engineering Society, May 1998.

[5] Takehiro Sugimoto, Yasushige Nakayama, and Satoshi Oode. “Bit Rate of 22.2 Multichannel Sound Signal Meeting Broadcast Quality”. In: *AES 137th Convention in Los Angeles, CA, USA*. Audio Engineering Society, Oct. 2014.

[6] Denis Martin et al. “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, Oct. 2016.

[7] ITU. “Methods for the subjective assessment of small impairments in audio systems”. In: *Recommendation ITU-R BS.1116-3* (2015).

[8] Josh Coalson. https://xiph.org/flac/comparison.html. 2018.

[9] Sascha Dick, Nadja Schinkel-Bielefeld, and Sascha Disch. “Generation and Evaluation of Isolated Audio Coding Artifacts”. In: *AES 143rd Convetion in New York, NY, USA*. Audio Engineering Society, Oct. 2017.

[10] Chi-Min Liu et al. “Compression Artifacts in Perceptual Audio Coding”. In: *AES 121st Convetion in San Francisco, CA, USA*. Audio Engineering Society, Oct. 2006.

[11] ITU. “Algorithms to measure audio programme loudness and true-peak audio level”. In: *Recommendation BS.1770-4* (2015).