Perceived Audio Quality of Realistic FM and DAB+ Radio Broadcasting Systems

JAN BERG,[1] *AES Member,* **CHRISTOFER BUSTAD**[2]*,*
(jan.berg@ltu.se,) (christofer.bustad@sr.se,)

LARS JONSSON,[2] *AES Member,* **LARS MOSSBERG**[2]*,* **AND DAN NYBERG,**[1,3] *AES Member*
(lars.jonsson@sr.se,) (lars.mossberg@sr.se) (dan.nyberg@skl.police.se)

[1]*Luleå University of Technology, Piteå, Sweden*
[2]*Swedish Radio, Stockholm, Sweden*
[3]*Swedish National Laboratory of Forensic Science – SKL, Linköping, Sweden*

The perceived audio quality of a digital broadcasting system such as DAB+ is dependent on what type of coding and bit rates are applied. Due to bandwidth constraints, audio quality is prone to be in conflict with other service demands such as the number of channels and the transfer of ancillary data. Compared to DAB+, several other audio services have superior bit rates that challenge the audio quality of DAB+. This paper reviews audio-quality criteria and investigates how the perceived audio quality of different broadcasting systems complies with the criteria. Two studies were conducted: Test 1 where DAB+ codecs were tested at bit rates between 96 and 192 kbit/s and Test 2 where DAB+ systems between 48 and 192 kbit/s as well as FM systems were tested. The systems in Test 2 were designed to as far as possible model a realistic broadcasting signal chain including commonly used dynamic processors. The studies were realized through two listening experiments using the ITU-R recommendations BS.1116 and BS.1534 (MUSHRA) followed by a closing interview. The results showed that the currently highest available subchannel bit rate for DAB+ (192 kbit/s) was insufficient for attaining perceptually transparent quality for critical items, whereas it enabled a quality comparable to or in some instances better than a modern FM system. Extrapolation of data indicates that critical items may need to be coded at even higher bit rates to reach perceptually transparent quality. From the interviews, auditory features important for the subjects' assessment of quality were observed. This study concludes that when making decisions on broadcasting systems, it is important to have well-founded and clearly defined criteria for minimum acceptable quality and/or perceptually transparent quality.

1 INTRODUCTION

A broadcasting system generally consists of different circuits for contribution, distribution, and emission. These parts refer to different the following functions: contribution – the network between production sites; distribution – the network delivering the programme to the transmitter; and emission – the radio frequency (RF) signal. The signal processing (e.g., by codecs) taking place in each of these circuits may be of different types that utilize different bit rates depending on the system design, and as a result, it can have an impact on the audio quality [1]. It is, therefore, important to study how different codecs and bandwidths used in broadcasting affect the perceived audio quality. It is also important to define a minimum level of audio quality in order to have a yardstick that the quality of systems may be measured against. Knowledge about these relations would be valuable for assessing today's broadcasting systems as well as for designing future ones.

In contribution circuits, a significant number of steps of re-encoding must be made in the contribution chain during the production of programme content, for example, in editing or other manipulation of the programme [2]. In this way, a number of codecs are cascaded and as a result, cascading artifacts are added to the final programme before distribution. To avoid such artifacts, the bit rate in cascaded systems has to be increased [3]–[5]. This is also the reason for why Audio over IP (AoIP) contribution circuits require higher bit rates and thus bandwidths than those used in systems for distribution and emission. The available bandwidths have increased over time and will continue doing so [6]. The increase in the actual bit rates of IP networks takes place at a more or less unchanged price, which actually allows the use of sufficiently high bit rates to avoid the

Table 1. Examples of codecs and typical bit rates used in current consumer systems and in DAB and DAB+.

System	Examples of codecs	Typically used bit rates [kbit/s]
DAB+	AAC/HE-AAC	32 . . . 128
DAB	MPEG-1/2 Layer II	64 . . . 192
DVB	MPEG Layer II Dolby Digital	192 . . . 256 (stereo)**, 448 (stereo + multichannel)
Blu-ray disks	PCM*/Lossless coding	$\geq$ 6 Mbit/s (stereo + multichannel)
DVD	PCM*/DTS/Dolby Digital	2304 (stereo), 640 . . . 1500 (multichannel)
Online music catalogs	FLAC*	$\geq$ 800 (stereo)
Web streaming	MPEG-1 Layer III/Windows Media Audio/AAC	32 . . . 320, 128 typical
iTunes	AAC	128 . . . 256
Spotify	Ogg Vorbis	96 . . . 320
Wimp	AAC/HE-AAC	64 . . . 256

* In PCM (Pulse Code Modulation) and FLAC (Free Lossless Audio Codec) no lossy data compression has been used.
** Bit rates used in Sweden.

cascading artifacts. Today, there is no reason to keep the very low bit rates that were necessary when AoIP contribution systems were introduced just after 2000 [7]. Although the need for increased bit rates has been reported, these results are not widely known so it has not influenced the roll out of bit rate compressed audio. More discussion, research, and listening tests on the cascading performance of, for instance, Advanced Audio Coding (AAC) [8] are needed to understand the phenomenon even better.

When it comes to distribution and emission, broadcasters often have heated and unresolved debates about the best bit rates to use when digitally broadcasting via Web radio and over Digital Video Broadcasting (DVB) or digital radio airwaves. When designing a digital service, there are two contradicting targets both of which are meant to fit within the investment level that has been determined. One target is using high enough bit rates so perceived audio quality is deemed acceptable; the other target is the capacity to transmit as many channels as possible. The total cost of ownership of networks and transmitters often plays a dominant role for selection of bit rates, limiting the upper level of audio quality. Consequently, striking an appropriate balance between the number of channels and their audio quality is a delicate and crucial decision. In broadcasting, the relationship between perceived audio quality and bit rates is continuously being evaluated and discussed within and between radio and television companies and research bodies [4],[5],[9]–[11]. In some cases, the selected bit rates are determined simply by just testing their impact on the public during a real broadcast; alternatively, bit rates are sometimes based on listening test results. The answers to questions about what audio quality is desired and acceptable, how this relates to descriptors such as "good enough," "good," and "transparent" and what these actually correspond to in terms of bit rates are crucial for making an informed decision about the requirements of a broadcasting system.

Distribution or emission of audio over the airwaves is under pressure to maintain a low bit rate in order to cope with the ever-more crowded and expensive broadcast radio spectrum. In some cases, broadcasters are challenged by the wireless industry to give up some parts of their frequencies [12]. Systems using 3G and 4G LTE (Long-Term Evolution) [13] as well as Wireless Local Area Network [14] via broadband now provide more bandwidth for audio transfer, for example, via smartphones or different systems for home listening.

Earlier, due to the bandwidth limitations, the general public had been forced to listen to quite low bit rates such as those in DAB in European countries and MP3 downloads used by portable devices [15]. In the light of the changes in availability of bandwidth, the question whether audio quality levels are too low for the distribution and emission of DVB audio and digital radio should now be addressed. A part of this discussion includes examining whether it is acceptable that this level of quality is lower than the quality of other popular consumer systems, such as Compact Disc (CD), Digital Versatile Disc (DVD), Blu-ray, and FM radio. In addition to physical media, downloadable high-quality counterparts[1] aimed for different reproduction formats (e.g., surround sound in 5.1 configuration) exist as well as streaming applications such as iTunes[2], Spotify[3], and other formats [16] that also provide listeners with higher bit rate audio. Examples of different systems and the associated codecs and typical bit rates are found in Table 1. Clearly, today typical DVB and consumer systems use higher bit rates than digital radio broadcasting systems are capable of, for example, DAB+ maximum bit rate is 192 kbit/s, whereas DVB and iTunes allow for higher bit rates.

When discussing audio quality, several quality criteria occur; the concept of "transparency" is particularly important, although it can be interpreted in several ways [17]. One common use is "bit transparent." Bit transparent refers to when a Pulse Code Modulation (PCM) sample passes through some apparatus unaffected with identical content on input and output bit-by-bit [18]. Another transparency definition that describes perceptually transparent quality might be called "perceptual transparency"; this is also described as "audible transparent" or "transparent to the human listener" [17]. Perceptual transparency means people cannot perceive any changes in audio quality when comparing processed audio with the unprocessed original. To

[1] E.g. www.hdtracks.com/, www.rdio.com, www.wimpmusic.se
[2] www.apple.com/itunes/
[3] www.spotify.com/

assess perceived quality, listening tests are required [19] and frequently used assessment methods are the ITU-R recommendations BS.1116 (for small impairments) [20] and BS.1534, commonly referred to as MUSHRA (for intermediate quality) [21]. In these methods, subjects assess the Basic Audio Quality (BAQ) of signals in the form of sound excerpts that have been processed by codecs. The processed excerpts are commonly referred to as items and they are graded in relation to the unprocessed excerpts that form reference signals. In a listening test, perceptual transparency means that the audio quality score of a perceptually transparent item should not show any statistically significant difference from the score of the reference signal.

A second criterion, "acceptable" broadcast audio quality, requires a better score for all items than one grade down on the 5-grade evaluation scale used in ITU-R BS.1116, that is, a Subjective Difference Grade (SDG) > -1.0. This limit is the lowest allowed result of a listening test for production or contribution circuits. In European Broadcasting Union (EBU) Tech 3339, this interpretation is described in the following way:

> If cascaded codec chains are to be considered for broadcast use then the quality criterion should be that none of material should produce an average diff-grade worse than -1.0 ("*perceptible but not annoying*"). If all the tested items score better than -1.0, then we can consider the chain's performance to be acceptable. However, if the average over all items is better than -1.0, but some test items score worse than this, then we must be wary of using such a chain. If the average over all items is worse than -1.0 then the chain should not be considered acceptable for broadcast use. [5]

As this quotation suggests, even when a specific codec shows an overall acceptable performance averaged across items, certain single items may still show an unacceptable quality, a concern that any analysis should consider. Similarly, EBU BPN 019 [9] uses the phrase "indistinguishable quality" and EBU BPN 094 [10] uses the phrase "broadcast quality" to define acceptable quality. The latter notion seems to be implicitly defined as excellent quality or a score > 80 on the MUSHRA 100-point scale [21] in EBU BPN 094, Section 9.1.

In addition, there is a possible third audio quality criterion, "FM quality." As FM has been the predominant way of audio broadcasting, it is a de-facto reference point, or a sort of anchor that later systems can and will be compared with by listeners. Consequently, a comparison between DAB+ and FM is of interest. Such a comparison would not be straightforward as the systems under test will be susceptible to different forms of quality degradation under real-life conditions. However, if conditions are specified, a comparison is possible.

Clearly, several possible definitions of audio quality of a system exist and a wide span of bit rates, many of them lower than what is employed by other audio applications, are used for DAB+. The complete signal path of digital broadcasting includes multiple audio coding and decoding points where signal degradations occur. To handle these issues for future audio broadcasting, it will be important to understand which quality criteria are suitable and how these should be interpreted and implemented.

If the bit transparency criterion is omitted, the remaining criteria discussed builds on perceived audio quality measures resulting from listening tests. These audio-quality criteria will be the focus of the current study:

1. *perceptual transparency*, which means that no statistically significant difference between the reference and a tested system should be found;
2. *broadcast quality*, which means that the mean Basic Audio Quality of a system across all items, and preferably also the mean of each item, should fulfill SDG > -1.0 in BS.1116 or Score > 80 in MUSHRA; and
3. *FM quality*, which means that the tested system should show equal quality, for example, no statistically significant difference, when compared with a specific FM system.

This paper investigates the perceived audio quality of FM and DAB+ systems at different bit rates and how the perceived audio quality compares with quality criteria. In one part of the study, dynamic processors that would be commonly encountered in professional audio broadcasting are inserted into the audio path before encoding. Such a path is henceforth referred to by the term "realistic system." Whether the quality provided by the systems is sufficient is also discussed. This study has five objectives: (i) to assess the perceived audio quality of higher bit rates in DAB+ (4 bit rates, 96 ... 192 kbit/s); (ii) to assess the perceived audio quality of both realistic FM systems (2 configurations) as well as of low and high bit rate realistic DAB+ systems (6 bit rates, 48 ... 192 kbit/s); (iii) to investigate what additional information about the perceived audio quality of such systems can be elicited by conducting interviews after listening tests; (iv) to test the compliance of the results with the quality criteria 1, 2, and 3 above; and (v) to estimate what bit rates are likely to be required for perceptual transparency of the systems under test.

Section 2 contains method and results from listening tests and interviews whereas the tested systems' compliance with the audio quality criteria is treated in Section 3. The findings are discussed in Section 4.

2 LISTENING TESTS

In the following experiments, a number of coding processes possible for broadcasting purposes will be investigated for their perceived audio quality. Two test methodologies were used in these experiments, which were divided into Test 1 and Test 2. The recommendations used were in Test 1 the ITU-R BS.1116 [20] and in Test 2 the ITU-R BS.1534 (MUSHRA) [21]. In both tests, Basic Audio Quality (BAQ) was assessed by the same group of subjects. BS.1116 was chosen to detect expected small differences between the systems tested in Test 1, whereas MUSHRA was used for the expected bigger quality differences due to

Table 2. Codec settings in Test 1.

Name	Subchannel bit rate [kbit/s]	Audio bit rate (including PAD) [kbit/s]	PAD bit rate [kbit/s]	SBR	Parametric stereo (PS)
DAB+ 96 (SBR)	96	87.2	1.33	Yes	No
DAB+ 128	128	115.8	2.53	No	No
DAB+ 160	160	145.1	2.53	No	No
DAB+ 192	192	174.5	2.53	No	No

the inclusion of low bit rate systems in Test 2. Test 1 was always performed first by the subjects due to the expected smaller differences between codecs in Test 1, thus avoiding a harder task at the end when listener fatigue could influence the results. Possible order effects between the two tests were not investigated. Each subject first trained for Test 1 (BS.1116) before performing Test 1 and trained for Test 2 (MUSHRA) before performing Test 2. Normally, a break with coffee or tea and sandwiches was provided after the training for Test 2.

The general criteria for selection of audio excerpts to be included in the tests can be summarized as follows:

- The excerpts should span a broad range of different types of material and musical genres.
- The excerpts should come from typical programme material.
- The excerpts should clearly reveal something about the performance of one or more of the systems and of differences between the systems.
- There must not be any bias towards or away from any particular system.
- The excerpts should both be material previously used in other listening tests (to compare with other tests) and material not previously used in other listening tests (to avoid the possibility that codecs might be tuned to the excerpts selected for the test).
- The excerpts should not be wearisome or too involving.

Interviews were conducted after the listening tests in order to identify the subjects' experience in terms of what degradations the subjects perceived and how the degradations were perceived. In this paper a subset of interviews was randomly selected for analysis.

Loudness levels in LUFS refer to recommendation EBU R 128 [22].

2.1 Test 1

2.1.1 Coding processes

All audio selected for the test was encoded into the specific type of High-Efficiency (HE) AAC used in DAB+ using a command line software encoder called "testenc" version 1.2.0 (build July 16, 2007) from Dolby and then decoded using the Dolby command line decoder "testdec" version 1.0.0 (build 31 May 2007). The codecs were previously used in a DAB+ test, EBU D/DABA project report BPN 094 [10].

In DAB+, the bit rate for a subchannel is not only used for audio data but also for programme-associated data (PAD), error protection, etc. The bit rate available for audio data is about 88–91% of the subchannel bit rate (due to error protection) minus the bit rate used for PAD [23]. The highest subchannel bit rate allowed in DAB+ is 192 kbit/s and to keep the number of different bit rates at a level that would avoid a too lengthy test and possible listener fatigue, four bit rates spaced 32 kbit/s apart were chosen. Consequently, the subchannel bit rates used by the encoder was 96, 128, 160, and 192 kbit/s. Spectral band replication (SBR) was only used for the lowest bit rate.

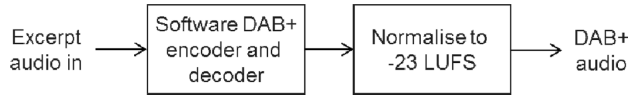


Fig. 1. Block diagram of Test 1.

The bit rates for PAD were 1.33 kbit/s for the lowest bit rate with SBR and 2.53 kbit/s for the three higher bit rates without SBR. These values were chosen to be the same as in the listening test reported in EBU BPN 094 [10]. There are, however, a number of services for which higher PAD bit rates probably would be selected and the maximum bit rate for PAD can be set as high as 79 kbit/s for 48 kHz sample rate when SBR is disabled. If a higher PAD bit rate is selected and if the subchannel bit rate remains the same, the audio quality will decrease. In this experiment, the audio bit rates including PAD were 87.2, 115.8, 145.1, and 174.5 kbit/s. For more details of the PAD insertion, see Section 7 in EBU BPN 094. The sampling frequency was always set to 48 kHz and all audio was encoded in stereo. See Table 2 for names of the settings and an overview of the bit rates.

One excerpt in this listening test (Speech (pan)) was encoded and decoded in two versions. The first version was the original speech signal and the second version was the same audio sent through an MPEG-1 Audio Layer II encoder at 384 kbit/s, stereo, 48 kHz, and a decoder five times. The encoder and decoder was a special version of Awave Audio by FMJ Software called Awave SR version 2.3, which internally uses TooLame. The audio bandwidth of this encoder at this bit rate exceeded 20 kHz. As previously discussed, the production at broadcasting companies often includes a number of cascades, for example, for Swedish Radio this is valid for MPEG-1 Audio Layer II. In order to investigate its influence on the audio quality the cascade was included.

Each item to be used in the listening test was finally adjusted to have a loudness of −23 LUFS and was also carefully synchronized with other items originating from the same excerpt. See Fig. 1 for a simplified graph of Test 1.

Table 3. Excerpts used in Test 1.

Name of excerpt	Name in BPN 094	Description	Length [s]	Upper frequency limit [kHz]
Applause w announcer	g_applause, and female announcer	Applause with female announcer.	16.5	18.3
Classical	f_brass, timpani and castanets	Brass, timpani and castanets, from Manuel de Falla's ballet *El Sombrero de tres picos*. Taken from track 1 on the SACD HMC 801606 from Harmonia Mundi.	17.8	24
House	a_electro pop	Excerpt from the house/pop song "Love is Gone" by David Guetta.	20.1	22
Speech (pan)	b_female speech Swedish	Female speech from a Swedish newscast panned slightly to the right. (Inter-channel level difference = 6 dB.)	16.0	24

2.1.2 Subjects

All the subjects ($N = 30$) had experience in listening critically to reproduced sound and did therefore meet the basic requirements of BS.1116. They were recruited at two locations in Sweden. In Stockholm, the subjects ($n = 17$; aged 21–62; mean = 41 years; median = 41) included personnel from Swedish Radio as well as independent listeners although most listeners were professional sound engineers. At the second location, the School of Music at Piteå, a campus at Luleå University of Technology, the subjects ($n = 13$; aged 19–29; mean = 23; median = 21) all had audio technology education and were either students or alumni at the school. Six of the subjects had undergone hearing tests in their admission tests for their educational program. The remaining subjects were assumed to have regular hearing (none of them stated a known hearing loss before the test).

2.1.3 Stimuli

As stated above in the criterion for selection of excerpts, the excerpts should both be material previously used in other listening tests and material not previously used in other listening tests. Given that the maximum bitrate for DAB+ was to be used, four of the most critical excerpts used in BPN 094 were selected for this test. Since the same encoder used for BPN 094 was used for this test, there was no particular need to seek new excerpts. The four excerpts selected are found in Table 3. The excerpt "Speech (pan) + 5xL2" is the excerpt "Speech (pan)" after five encodings and decodings of MPEG-1 Audio Layer II as described in Section 2.2.1.3.

2.1.4 Equipment, listening environment and listening levels

The listening equipment was selected to match the previous BPN 094 test [10],[24]. Headphones were used to reduce the possible differences between room characteristics of the two test sites. The headphone amplifier Grace Design m903 contained a D to A converter using an accurate internal clock. Diffuse-field EQ was employed in the headphones (Sennheiser HD-650) to conform to the reference response in ITU-R BS.708 [24]. The filter characteristics were developed by the IRT for BPN 094 and were adapted for the HD-650 headphones specifically. The filter characteristics were employed in these tests by processing the audio files using convolution at a high resolution [24]. This amplifier was connected to a computer equipped with the STEP software[4], version 1.08a, via a USB interface. Jitter performance was measured in the analog domain with the Audio Precision System Two jitter signal. The values were found to be in the same order of magnitude or better than other high-performance converters.

The noise levels of the listening rooms were well under NR 15 at Piteå and under NR 20 in Stockholm.

According to reports from subjects in earlier listening tests, a fixed listening level did cause annoyance when it was perceived as being too low or to high for the individual subject. Reports also indicate that subjects have been distracted when the listening level at the training before the test did not match the level of the actual test [subjects in listening test, personal communication, 2011]. In order to overcome such possible problems in the current test, the listeners were free to adjust the listening level during the experiment. They were also instructed to make notes on what listening level they choose. After the postscreening process (Section 2.1.7), the remaining data showed that 52% of the listeners did not adjust their chosen level at all during the experiment. 62% did their adjustments within 3 dB and 76% kept the level variation within 6 dB. Consequently, the loudness alignment performed before the experiment seemed to remove most of the possible loudness differences.

2.1.5 Training

Subjects invited to the listening test were sent a link to audio files that they were asked to listen to at home for training purposes before the tests. The audio files contained all audio that the subjects were to grade in the actual tests. Everything that was not the original audio was relabeled as

[4] www.audioresearchlabs.com

random numbers so as not to give any clues of the systems they had passed through. Every subject received and was asked to use two randomized playlists for the audio files. If some subjects used playback equipment of varying quality or choose not to listen to all audio files, this procedure would distribute the training sessions randomly over all audio files.

The actual test was preceded by training that consisted of grading a random subset of the audio that was to be graded later in the actual test. This way each subject learned the functions of the user interface, became familiar with the listening environment and equipment, and practiced grading some of the audio from the actual test. The training consisted of a subset comprising five gradings, randomly selected for each subject. Each of the five excerpts appeared once, three of the four bit rates appeared once, and one bit rate appeared twice.

2.1.6 Procedure

In each trial, signals were presented pairwise by means of the graphical user interface (GUI), one signal always being the original unprocessed signal (the hidden reference). The second signal was a processed (coded and decoded) version of the reference signal. The two signals were randomly assigned to playback buttons labeled "A" and "B." In addition, the reference signal was available from a separate playback button marked "REF." The subjects' task was to grade the Basic Audio Quality of both A and B compared with REF on the 5-point scale prescribed in recommendation BS.1116. The scale goes from 1.0 through 5.0, accurate to one decimal place. An imperceptible difference between the signal under assessment (A or B) and REF should be indicated by assigning 5 to that signal. A signal exhibiting any perceived difference from REF should be indicated by a grade < 5. At least one of the two signals has to be assigned grade $= 5$, as one of them is identical to the reference. The subjects recorded their assessments through the GUI by pulling grading sliders to a scale position that corresponded to their judgment. For every signal pair (reference and processed signal), the procedure was performed in random order for all combinations of codecs and excerpts, yielding a total of 20 trials. The GUI including scale labels, grading sliders and playback controls is depicted in Fig. 2.

Two test leaders worked together during the test. At the time of the tests, each subject received the same instructions and everyone was instructed by the same two test leaders. In almost all cases, both test leaders were present. In this way, the two test leaders complemented each other to ensure that all subjects received very similar oral instructions before the test. The test leaders used a checklist to make the instructions as consistent as possible. The test was totally blind to the listeners. The test leaders were careful not to influence the subjects in how they graded different artifacts and the type of systems or codecs tested were never mentioned.

The following information and instructions were given to the listeners before Test 1:

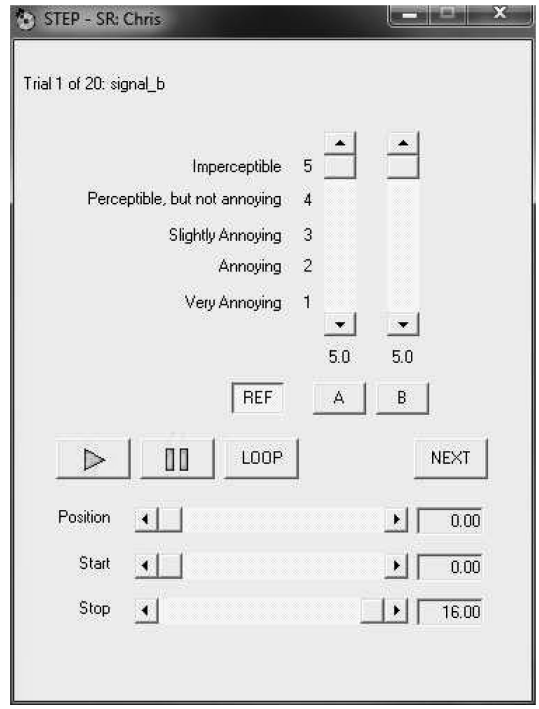


Fig. 2. Graphical user interface used in Test 1.

1. A general background to the test.
2. An overview of the test methodology and user interface.
3. To take regular breaks.
4. Usage of the looping functionality was only allowed when a subject had listened to all audio from beginning to end.
5. Subjects were informed about the level equalization applied. They were encouraged to set a comfortable listening level and were also allowed to adjust the listening level at any time during the test.
6. The audio quality change as a whole – for example, "Basic Audio Quality" (BAQ) – should be considered when deciding on a grade, which means that all audio-quality differences from the reference should be considered to be audio quality degradations.
7. Only the degradations in audio quality should be considered when deciding on a grade and not how good the mix is or how the listener enjoys the content.
8. The listener should be careful not to pull the wrong slider when giving a grade.
9. It is not possible to go back and change a previously given grade.
10. Listeners should make a note of the listening levels they used.

For each trial, a Subjective Difference Grade (SDG) was calculated by subtracting the grade assigned to reference from the grade assigned to the object under test. Thus, in a trial where the subject correctly identified the reference signal by assigning a lower grade to the processed signal than to the hidden reference signal, SDG < 0.

Table 4. ANOVA Table.

Source	Type III sum of squares	*df*	Mean square	*F*	Sig.	η_P^2
Corrected Model	421.883[a]	39	10.818	15.619	.000	.616
Intercept	991.737	1	991.737	1431.943	.000	.790
Codec	117.840	3	39.280	56.715	.000	.309
Excerpt	253.043	4	63.261	91.340	.000	.490
TestSite	7.724	1	7.724	11.152	.001	.029
Codec * Excerpt	12.097	12	1.008	1.456	.139	.044
Codec * TestSite	2.371	3	.790	1.141	.332	.009
Excerpt * TestSite	3.862	4	.965	1.394	.235	.014
Codec * Excerpt * TestSite	8.175	12	.681	.984	.464	.030
Error	263.181	380	.693			
Total	1723.150	420				
Corrected Total	685.064	419				

a. *R* Squared = 0.616 (Adjusted *R* Squared = 0.576)

2.1.7 Postscreening

Postscreening identifies and rejects subjects that show an inability to discriminate the reference (unprocessed) signals. The procedure follows the recommendations in BS.1116. First, any item that received low average grade across all subjects, that is, having a mean SDG below −2.0, was temporarily removed during the following stage. Second, for every subject, the remaining data thus obtained was subjected to a one-sided *t* test ($\alpha = 0.05$) to assess the likelihood that the mean of the distribution for each subject is zero or greater. Subjects failing to produce a mean that is significantly less than zero at $\alpha = 0.05$ were rejected from the subsequent analysis. As a result of this process, nine subjects were rejected and the associated data were removed from the data set. Hence, data from 21 subjects remained for analysis. After this adjustment, the data set contained 420 data points (21 subjects * 4 codecs * 5 excerpts). Both test sites were equally represented among rejected subjects. The median age was for rejected subjects = 39 and for nonrejected subjects = 29. The age difference may indicate that older subjects found it more difficult to discriminate the type of differences in audio quality occurring in this test.

2.1.8 Data analysis

In Test 1, the experimental factors' effects on Basic Audio Quality represented by the dependent variable SDG were investigated by means of analysis of variance (ANOVA) after checking the assumptions underlying ANOVA (details in Section 2.1.9). The factors in the model were as follows: Codec = Coding process under test (4 levels, Table 2); Excerpt = Sound excerpt (5 levels, Table 3); and TestSite = Site where test was performed (2 levels, Piteå/Stockholm). The ANOVA model was determined using the following equation: SDG = Mean + Codec + Excerpt + TestSite + Codec*Excerpt + Codec*TestSite + Excerpt*TestSite + Codec*Excerpt*TestSite + Error. Post hoc tests were also made to investigate differences within the factors Codec and Excerpt. Additionally, the acquired data was used to find a model for BAQ as a function of bit rate, assuming a linear relation between those (see Section 3.3).

2.1.9 ANOVA

The ANOVA residuals were tested for normal distribution using the Kolmogorov-Smirnov test. As $K = 0.038$, ($p = 0.166$), normal distribution of residuals could not be rejected and is, therefore, assumed. In addition, for each combination of codec and excerpt, the data were checked for normal distribution by means of the Shapiro-Wilk test. Out of the 20 combinations, the null hypothesis was rejected for five cases only ($\alpha = 0.05$).

This implies that the vast majority of the data come from a normally distributed population. Levene's test of equality of error variances yielded $F(39,380) = 1.227$, ($p = 0.172$). Hence, equal error variances could not be rejected and were, therefore, assumed. The experimental design included a randomization of both trial order and assignment of stimuli to the graphical user interface buttons within each trial. Thus independency between data points was ascertained. In summary, the assumptions underlying ANOVA were not violated. The ANOVA including the effect size, partial eta squared (η_P^2), is summarized in Table 4.

The ANOVA showed that all three main factors were significant. The largest effect size was found for Codec ($\eta_P^2 = 0.31$) and Excerpt ($\eta_P^2 = 0.49$). For the remaining factors and combinations, the effects were negligible ($\eta_P^2 < 0.05$), so the factors Codec and Excerpt were further investigated and subjected to post-hoc tests.

2.1.10 Multiple comparisons of codecs

The BAQ (mean SDG) for different codecs across all excerpts are shown in Fig. 3. A Tukey HSD post-hoc test ($\alpha = 0.05$) was performed to find the significant differences between codecs (Table 5). The results showed that there is a significant difference between the BAQ for every pairwise combination of the codecs. The BAQ rises significantly for each increase in bit rate. Although the DAB+ codec at 192 kbit/s received the greatest SDG, its quality is still inferior compared with the reference. Comparisons with the audio quality criteria will be discussed in Section 3.

2.1.11 Multiple comparison of excerpts

The BAQ for different codecs and excerpts are shown in Fig. 4. A Tukey HSD post-hoc test ($\alpha = 0.05$) was

Table 5. Multiple comparisons of codecs (Tukey HSD); mean Basic Audio Quality [SDG] and resulting homogenous codec subsets (CS).

		Codec Subset			
Codec	*N*	CS1	CS2	CS3	CS4
96 kbit/s (AAC+SBR)	105	−2.303			
128 kbit/s (AAC)	105		−1.820		
160 kbit/s (AAC)	105			−1.313	
192 kbit/s (AAC)	105				−0.852
Sig.		1.000	1.000	1.000	1.000

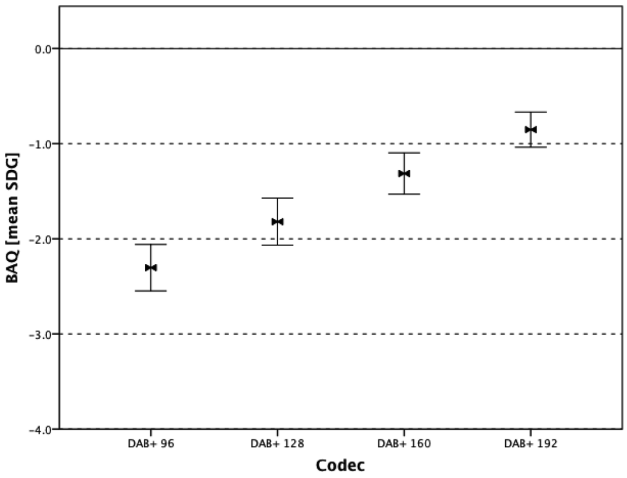


Fig. 3. Mean Basic Audio Quality and 95% confidence intervals for codecs across excerpts.

Table 6. Multiple comparisons of excerpts (Tukey HSD); mean Basic Audio Quality [SDG] and resulting homogenous excerpt subsets (ES).

		Excerpt subset		
Excerpt	N	ES1	ES2	ES3
Speech (pan)	84	−2.554		
Speech (pan) + 5xL2	84	−2.496		
House	84		−1.190	
Applause w announcer	84			−0.830
Classical	84			−0.790
Sig.		0.992	1.000	0.998

performed to find the significant differences between excerpts across codecs (Table 6). The results showed that there were three excerpt subsets (ES) where the excerpts within each subset were not significantly different from one another, but differentiated significantly from the excerpts of the other subsets. For each subset, the content was as follows in the order of increasing BAQ: ES1 – Speech (pan) and Speech (pan) + 5xL2; ES2 – House; and ES3 – Applause w announcer and Classical.

2.2 Test 2

2.2.1 Coding processes

As this is common practice in today's broadcasting, FM and DAB+ in this experiment both use one band and multiband audio processing to reduce unintentional level differences, to adapt the dynamics of the audio content for the listening environments common to listeners, to reduce

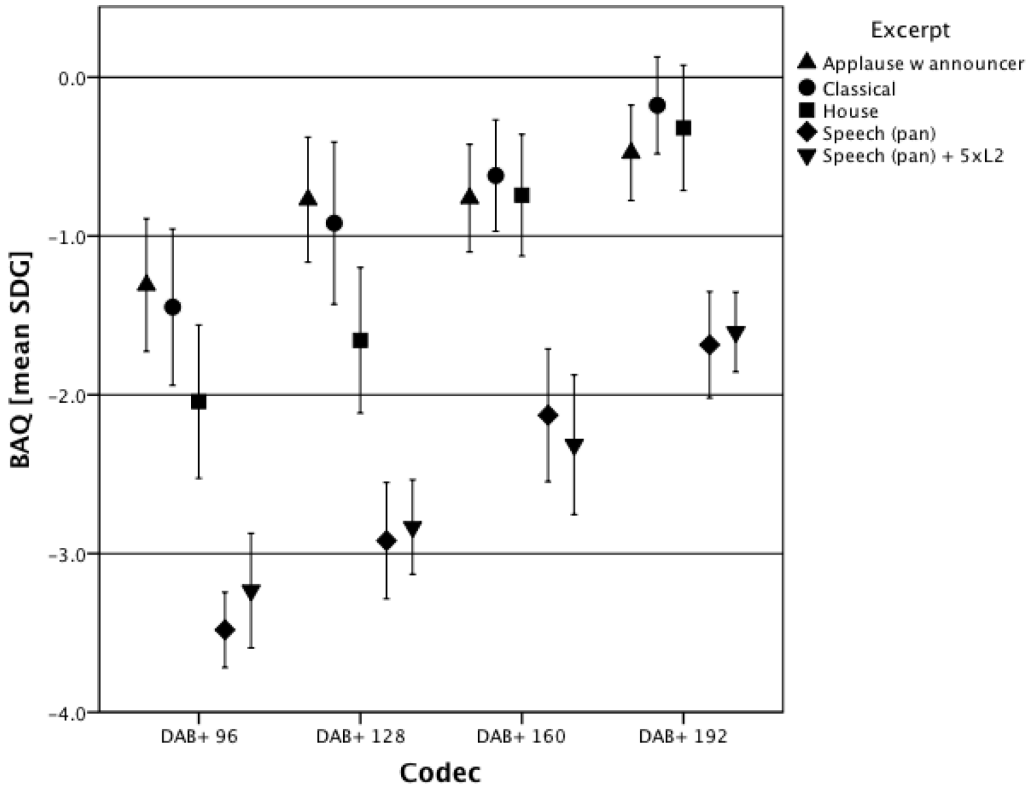


Fig. 4. Mean Basic Audio Quality and 95% confidence intervals for codecs and excerpts.

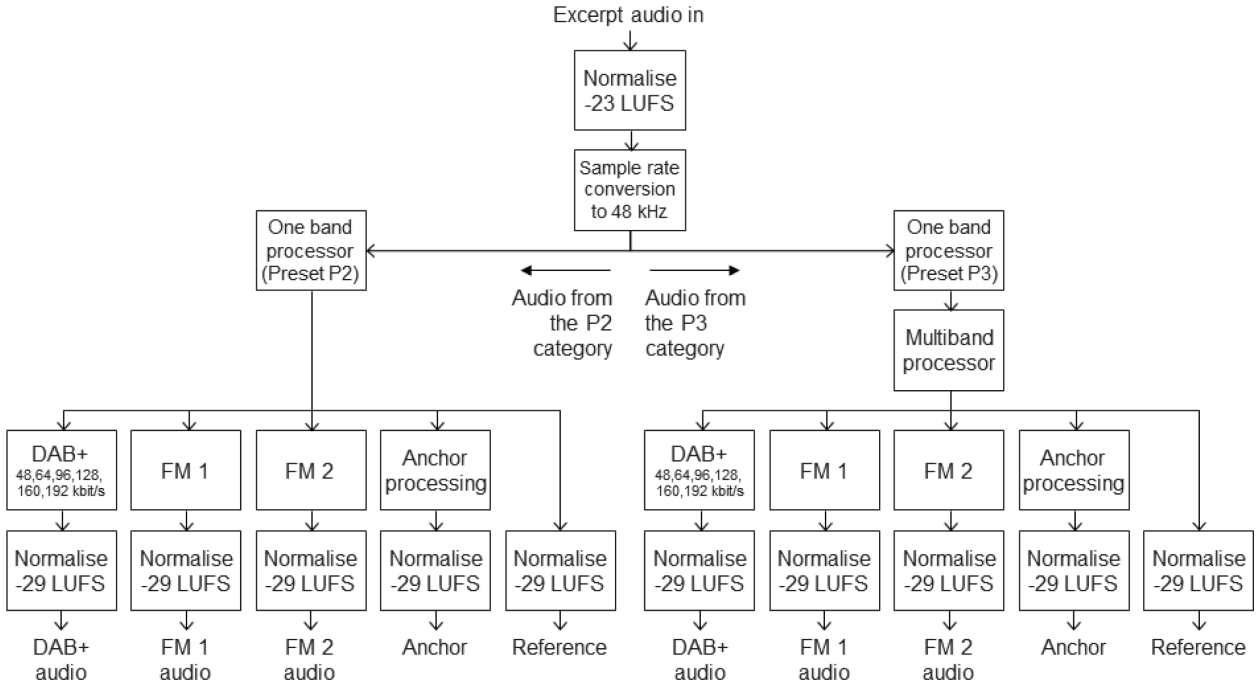


Fig. 5. Block diagram of Test 2.

the frequency response differences between different audio content, and, in some cases, to create a certain "sound" for a particular channel. In FM, one band (sometimes referred to as wide-band) and multiband processing is also usually closely adapted to the pre-emphasis limiting necessary for the FM stereo system. To create realistic broadcast systems, it was decided to include one band and multiband audio processing in the DAB+ and FM systems to be tested.

Since different audio content requires different types of audio processing, two processing categories were created. One category, P2, corresponded to the radio channel Swedish Radio P2, which broadcasts classical and contemporary music as well as jazz and folk music. The other category, P3, corresponded to the radio channel Swedish Radio P3, which broadcasts pop music, news, and cultural and social programmes. The processing for the P2 category was decided to include only one band audio processing and no multiband audio processing and the processing for the P3 category was decided to include one band processing followed by multiband audio processing.

To compare the audio quality of FM and DAB+ on the same terms, it was important for the audio processing to be the same for both systems, with the exception of the pre-emphasis limiting for FM. To ensure this consistency, audio content that later might be selected for the listening test was sorted into the two categories, P2 and P3. The total amount of audio in the two categories was about 7 h and was later reduced to a few excerpts. The loudness of each audio item in both categories was first normalized to −23 LUFS. Then audio in the P2 category was processed using the advanced one band audio processor Factum Cadenza[5] in such a way that the resulting loudness for everything in the category as a whole reached −23 LUFS. Using this strategy meant that the overall loudness corresponded to the recommended loudness according to EBU Tech 3344 [25]. The preset used, Preset P2, was a slight modification of the preset currently used by the Swedish Radio P2 channel during evenings. Preset P2 performed a minimum of short-term limiting and mainly raised the level of soft passages in the audio.

The audio in the P3 category was processed using the same one band audio processor followed by the software multiband audio processor Breakaway[6]. The preset used for the one band processor in this category, Preset P3, was a slight modification of the preset currently used by the Swedish Radio P2 channel during daytime. This preset also did a minimum of short-term limiting and mainly raised the level of soft passages in the audio, but more so than the Preset P2. The multiband processor used a gentle preset that did some dynamic compression with a moderate compression ratio in six frequency bands. The drive into its final limiter was set in such a way that the resulting loudness for everything in the category as a whole reached −23 LUFS.

The processed audio for the P2 and the P3 categories was fed into a DAB+ encoder and also into two different FM systems in which the broadcast processors only did pre-emphasis clipping and no other audio processing. The drives into the pre-emphasis clippers were set so that the MPX Power at some point reached but never exceeded the MPX Power limit defined in ITU-R BS.412 [26]. In this way, all the systems used audio processing that is realistic and normally occurring in broadcasting. The reference signals were extracted from the output of the audio processors

[5] www.factum.se

[6] www.claessonedwards.com

Table 7. Codec settings in Test 2.

Name	Subchannel bit rate [kbit/s]	Audio bit rate (including PAD) [kbit/s]	PAD bit rate [kbit/s]	SBR	Parametric stereo (PS)
DAB+ 48 (SBR, PS)	48	43.2	1.3	Yes	Yes
DAB+ 64 (SBR)	64	57.9	1.3	Yes	No
DAB+ 96 (SBR)	96	87.2	1.3	Yes	No
DAB+ 128	128	115.8	2.7	No	No
DAB+ 160	160	145.1	2.7	No	No
DAB+ 192	192	174.5	2.7	No	No

(multiband/one band). See Fig. 5 for a block diagram of Test 2.

2.2.1.1 DAB+ Given the present time restrictions, only a subset of the 7 h of processed audio was fed into an actual DAB+ encoder. All 7 h of audio were encoded into HE AAC using a command line encoder from Dolby and decoded back to PCM. This type of HE AAC, however, was not exactly the one used in the DAB+ system, but should approximate the correct type well. The bit rates used by the command line encoder were set to the actual audio bit rates used by the DAB+ encoder, that is, bits used in DAB+ for error correction and PAD were taken into account.

The selected excerpts for the listening test were fed into the DAB+ encoder Factum MAP250E and the audio from the monitor output was recorded. The audio was also sent over the air simultaneously through a DAB+ multiplex and a transmitter placed in the Nacka region of Stockholm, and the audio from a DAB+ receiver was recorded. As no differences could be found between the audio from the monitor output and the receiver, the audio from the monitor output was chosen for the listening test.

The subchannel bit rates used by the DAB+ encoder were 48, 64, 96, 128, 160, and 192 kbit/s. SBR was only used for the three lower bit rates, and parametric stereo (PS) was only used for the lowest bit rate. The bit rates for PAD were 1.3 kbit/s for the three lower subchannel bit rates and 2.7 kbit/s for the three higher subchannel bit rates. These values were chosen to be similar to those used in EBU BPN 094 [10] and in Test 1 in this paper. The audio bit rates including PAD were 43.2, 57.9, 87.2, 115.8, 145.1, and 174.5 kbit/s. The sample rate was always 48 kHz and all audio was encoded in stereo.

Each item to be used in the listening test was finally adjusted to have a loudness of −29 LUFS, not −23 LUFS as in Test 1. This adjustment prevented overshoots resulting from the audio codec from being clipped. In Test 2, one of the ten excerpts was a 22 dB attenuated copy of one of the other nine excerpts. Some of the items resulting from the attenuated version would be clipped in the normalization if they were normalized to −23 LUFS as in Test 1. Hence the loudness level −29 LUFS was chosen for Test 2.

Some items resulting from the nonattenuated version of the excerpt, however, were clipped in the HE AAC encoding or decoding. It is not known if the clipping occurred in the encoding and/or the decoding, since the effects were observed in the decoded audio. This clipping occurred at the bit rates 48, 64, and 96 kbit/s at some transients occurring in the excerpt. Since the audio level at the input to the encoder corresponds to the normalization standard EBU R 128, which in its turn corresponds to real broadcasting conditions, these items were kept and used in the listening test. Audio coding has been shown to produce overshoots up to 5.3 dB [27]. So far, unpublished research by one of the authors of this paper has shown overshoots up to 8 dB for 64 kbit/s. See Table 7 for DAB+ codec settings and an overview of the bit rates.

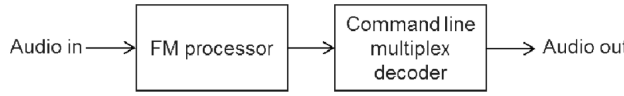


Fig. 6. Block diagram of FM 1.

2.2.1.2 FM The same processed audio that was fed into the HE AAC encoder and the DAB+ encoder was also fed into two FM systems. FM audio quality can mean many different things depending on the equipment used in the signal chain and in particular on the broadcast processor and how it is set to process the audio. For this reason, two FM systems were included in this listening test.

The first FM system, FM 1, consisted of a prototype of the audio processor Omnia 9[7] with 16 bit PCM input and output. Only its psychoacoustic multiplex clipper and a phase scrambler were active and no other processing was done in the processor. The phase scrambler could also be turned off. See section 2.2.3 for details on when the phase scrambler was in use. To attain an extended audio frequency range of about 17.5 kHz in this prototype, single sideband mode was active for the difference signals ($S = L - R$) in the FM multiplex. The audio processor also added an Radio Data System (RDS) signal with 6 kHz deviation.

In theory, an ideal FM modulation, demodulation, and reception would not affect the perceived audio quality. Hence, to get as close as possible to an ideal FM transmitter and receiver by removing the RF path, the multiplex output from the processor was saved directly to audio files at f_s = 192 kHz and decoded from multiplex to left/right audio using a command line utility[8]. To set the drive into the psychoacoustic multiplex clipper so that the multiplex power at some point reached but never exceeded the MPX Power limit defined in ITU-R BS.412, a command line utility was used that calculated the maximum of a moving average of the multiplex power. See Fig. 6 for a simplified graph of FM 1.

[7] omniaaudio.com

[8] Provided by Leif Claesson (see acknowledgments)

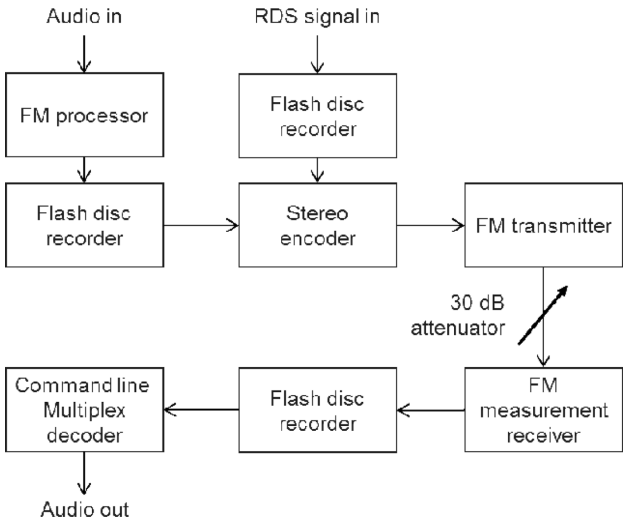


Fig. 7. Block diagram of FM 2.

The second FM system, FM 2, is of the same type as the most common FM system in Swedish Radio's FM network. This system's signal path started with the same audio processor as in FM 1 that here used a psychoacoustic left/right clipper and a phase scrambler and not its psychoacoustic multiplex clipper or any other processing. In this audio system, the phase scrambler could also be turned off. See Section 2.2.3 for details on when the phase scrambler was in use. De-emphasized left/right audio, limited to a 15 kHz bandwidth, and an RDS signal were fed into a Profline DMM stereo encoder with pre-emphasis enabled, and an analog multiplex signal from the stereo encoder was fed into a BW Broadcast TX5 FM transmitter. A 30 dB RF attenuator then reduced the output level from the transmitter into an Audemat FM-MC4 receiver (upgraded in software from FM-MC3). The multiplex output from this receiver was recorded on a SoundDevices 722 unit at 192 kHz mono 24 bits. Then the command line utility also used in FM 1 decoded the multiplex audio to left/right audio.

To set the drive into the psychoacoustic left/right clipper so that the multiplex power at some point reached but never exceeded the MPX Power limit defined in ITU-R BS.412, the same command line utility used in FM 1 was used on multiplexed audio before de-multiplexing and de-emphasis. The audio sent to the stereo encoder was left/right de-emphasized audio to imitate the current FM transmission chain for Swedish Radio. The audio was prepared in advance and played back on a SoundDevices 722 unit. The RDS signal sent to the stereo encoder was a band-pass filtered recording of an RDS signal played back on a SoundDevices 722 unit at 192 kHz sample rate. The deviation of the FM signal was set to 75 kHz, the RDS level was set to 3 kHz, and the signal strength into the receiver was set to 1 mV, all using measurements in the Audemat receiver. See Fig. 7 for a simplified graph of FM 2.

Each item to be used in the listening test from the two FM systems was finally adjusted to −29 LUFS to match the loudness level of the DAB+ systems. Each audio item was also carefully synchronized with other items originating from the same excerpt.

2.2.1.3 MPEG Audio Layer II It was discussed whether MPEG Audio Layer II as used in DAB should be included among the audio systems in this test. To avoid too many items to be graded by the subjects and since DAB using MPEG Audio Layer II will probably not be introduced in Sweden beyond trials, it was decided not to include MPEG Audio Layer II.

2.2.1.4 Anchor According to the MUSHRA recommendation, at least one anchor should be included for post-screening purposes and to make sure that items in the test cover a large portion of the audio quality scale. The required anchor in the recommendation is a 3.5 kHz low-pass filtered version of the reference signal. This anchor has been used in many tests over the years, but the anchor has lately been questioned as the audio quality of modern audio codecs has improved significantly. Since most codecs tested today have a much wider bandwidth and exhibit completely different artifacts compared to a 3.5 kHz low-pass filtered version, this anchor always stands out by being scored lowest, even if a test item has severe coding artifacts. This is the reasoning given by the D/DABA group in the EBU report BPN 094 to explain why a new kind of anchor that does not employ band limitation was developed and used. One example is found in EBU Tech 3324, Phase 1 and Phase 2 [11], where the confidence interval of the 3.5 kHz band limited anchor never overlapped the confidence interval of any of the other codecs in the test.

For the same reasons, the anchor developed by the D/DABA group was used in the MUSHRA test in this paper. This anchor is created by a simple algorithm that combines two types of signal impairment to reflect current typical coding artifacts. The first impairment is MDCT-based quantization distortion and the second is stereo image distortion. This algorithm is described in detail in BPN 094 and in an open source implementation[9]. The parameter settings used in this test were the same as used for the D/DABA tests (distortion set to 10 and fuzzy to 0.25) [10]. The anchor items were adjusted to −29 LUFS to match the loudness from the other items and the anchor items were also carefully synchronized with other items originating from the same excerpt.

2.2.2 Subjects

The subjects were identical to those participating in Test 1, see Section 2.1.2.

2.2.3 Stimuli

Audio content was first collected that was thought to be critical for one or more of the systems in the test. The 7 h of audio from the different categories and musical genres that were sorted into the two audio processing categories denoted "P2" and "P3" (as described in Section 2.2.1) were reduced using the criteria in Section 2. The audio processing

[9] sourceforge.net/projects/anchor

Table 8. Excerpts used in Test 2.

Name of excerpt	Preset	Description	Length [s]	Upper frequency limit [kHz]	Audio-processing category	Phase scrambler in FM systems	Commonalities with excerpts in Table 1 with the exception of the audio processing
Applause w announcer	P2	Applause with female announcer.	16.4	18.4	P2	On	Same as "Applause w announcer"
Classical	P2	Brass, timpani and castanets, from Manuel de Falla's ballet *El Sombrero de tres picos*. Taken from track 1 on the SACD HMC 801606 from Harmonia Mundi.	17.8	24	P2	On	Same as "Classical"
Electronic	P2	Excerpt with strong transients from "Vaihtovirta" on the album *Aaltopiiri* by Pan Sonic.	25.5	22	P2	On	
Electronic (att)	P2	Same as Electronic but first attenuated by 22 dB, i.e. the loudness is -51 LUFS, before encoding. (Later restored by 22 dB amplification after decoding.)	25.5	22	P2	Off	
House	P3	Excerpt from the house/pop song "Love is Gone" by David Guetta.	19.8	22	P3	On	Same as "House"
PopKent		The intro to the song "Dom Andra" by Kent.	19.6	21	P3	On	
PopRox	P3	A chorus in the song "Fading Like A Flower" by Roxette.	24.5	22	P3	On	
SpeechL (no pan)	P3	Female speech (long) from a Swedish newscast, centre panned.	26.4	24	P3	On	
SpeechL (pan)	P3	Same as SpeechL (no pan) above panned slightly to the right. (Inter-channel level difference = 6 dB.)	26.4	24	P3	On	A longer version of "Speech (pan)"
World	P2	Excerpt with handclaps and not much other percussion from the song "Migration" on the album *Introducing* by Nitin Sawhney.	21.4	22	P2	On	

for each category was performed and the processed audio was fed through the systems under test. The outputs from the different systems were then loudness aligned andsynchronized with the original processed audio.

Using the training mode in the software, it was possible to easily switch between the original processed audio and the outputs from the different systems, maintaining time synchronization. It was then noted which audio content was the most critical for the different systems. An overview of the 10 selected excerpts can be found in Table 8.

Preset "P2" indicates audio processing category P2; that is, that the audio processing was done using only an advanced one band audio processor that did a minimum of short term limiting and mainly low level compression. Preset "P3" indicates audio processing category P3; that is, the audio processing was done using the same one band audio processor as for the P2 category, again doing a minimum of short term limiting but more low level compression, followed by a software multiband audio processor using a preset designed to limit its influence on the audio signal. The audio processing for both categories are described in more detail in Section 2.1.1.

SpeechL, aside from the audio processing P3, is the same audio as Speech (pan) in Test 1 plus an additional 10.4 s of audio from the same newscast.

Electronic (att) is the same audio as Electronic but attenuated by 22 dB to make sure that no audio reached the clip level of the pre-emphasis clippers in the FM systems. This excerpt contains high-level transients at high frequencies, a condition that explains why the attenuation was chosen to be so high. After feeding this attenuated excerpt though the different systems, the level was restored through

amplification of the resulting items to match the loudness of the other items.

The phase scrambler in the FM audio processor makes asymmetric audio waveforms more symmetric and thus less sensitive to pre-emphasis clipping. The reason for its use is that certain types of signals such as speech can have higher peak levels on the positive side of the waveform than on the negative side or vice versa, and this asymmetry can be reduced by the phase scrambler. When more symmetry is reached, the highest peak levels are usually reduced. Experience shows that the use of the phase scrambler may cause quality alterations. Hence, the phase scrambler was enabled for all excerpts except Electronic (att) as this was attenuated. The excerpt Electronic (att) was designed not to trigger the pre-emphasis clipper and thus the phase scrambler was not necessary and therefore turned off.

Below is a list of features of the selected excerpts in Test 2:

- Five excerpts in the P2 audio processing category and five excerpts in the P3 category;
- Three excerpts containing modern mainstream music and four excerpts containing nonmainstream music;
- One excerpt containing applause;
- One excerpt containing panned speech;
- One excerpt containing non-panned speech;
- Two versions of an excerpt with strong transients with different levels to investigate how the pre-emphasis clipper influences the grading of the FM systems;
- Four excerpts that are, aside from the audio processing, the same as in Test 1 and in BPN 094; and
- Some excerpts containing a moderate amount high frequency energy and excerpts containing much high frequency energy.

All excerpts were edited to have smooth beginnings and ends and were calibrated to −29 LUFS so as to have the same loudness.

2.2.4 Equipment, listening environment, and listening level

The equipment, listening environment and listening level procedure were identical to that in Test 1, see Section 2.1.4. After the postscreening process (Section 2.2.7), the remaining data showed that 50% of the listeners did not adjust their chosen listening level at all during the experiment. Fifty-nine percent did their adjustments within 3 dB and 73% kept the level variation within 6 dB. Consequently, the loudness alignment performed before the experiment seemed to remove most of the possible loudness differences.

2.2.5 Training

A part of the training for Test 2 took place at the subjects' home. This followed the same outline as in Test 1 (Section 2.1.5). The on-site training for Test 2 (MUSHRA) consisted of a randomly selected subset of three excerpts processed in all versions. Each subset was randomly selected to distribute the training randomly over all subjects.

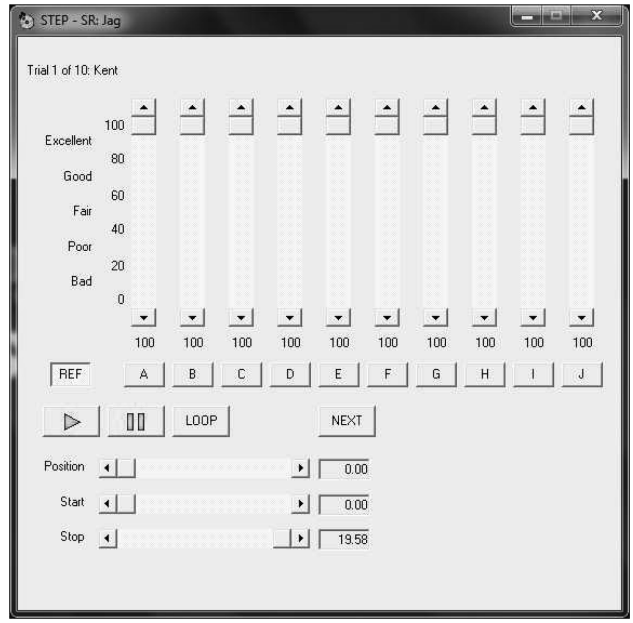


Fig. 8. Graphical user interface used in Test 2.

2.2.6 Procedure

In each trial, the Anchor signal and all eight processed versions of the excerpt and its original unprocessed signal (the hidden reference) were presented by means of the graphical user interface (GUI). In every trial, the signals were randomly assigned to 10 playback buttons labeled "A" through "J." In addition, the reference signal was available from a separate playback button marked "REF." The subjects' task was to grade the Basic Audio Quality of each signal (A ... J) compared with REF and with the other signals on the scale prescribed in MUSHRA. The integer scale goes from 0 through 100 and is divided into five parts of equal length where each part is associated with a descriptive label. An imperceptible difference between any of the signals under assessment (A ... J) and REF should be indicated by assigning 100 to the particular signal(s). A signal exhibiting any perceived difference from REF should be indicated by a score < 100. At least one of the signals has to be assigned Score $= 100$, as there is one hidden reference. The subjects recorded their assessments through the GUI by pulling grading sliders to a scale position that corresponded to their judgment. This procedure was performed for all excerpts in a random order, thus yielding 10 trials. The GUI including scale labels, grading sliders and playback controls is depicted in Fig. 8.

The following information and instructions were given to the listeners before Test 2:

1. The same instructions as for Test 1 (Section 2.1.6, points 1–10) apply, but this time the MUSHRA test methodology and user interface (described in the procedure above) should be used.

2. The audio in Test 2 is normalized to a lower level than for Test 1.

2.2.7 Post-screening

The postscreening was performed according to step 1 through 4 below and its objective was to reject subjects that were unable to discriminate between stimuli in terms of reference signals and anchor signals. The score range used by each subject was also investigated.

1. *Ability to identify reference signals*. The probability of correctly identifying the hidden reference signal by chance in one trial including ten signals is 0.1. The total number of trials is ten. The minimum required number of correct identifications for all trials, n_{ID}, at $\alpha = 0.01$ is calculated by means of the cumulative distribution function of the binomial distribution satisfying $\mathrm{b}(x; 10, 0.1) \geq 0.99$, where $n_{\mathrm{ID}} > x$. Consequently, a subject must show $n_{\mathrm{ID}} \geq 5$ to fulfill this condition. When applying this condition, seven subjects were rejected from the subsequent analyses.
2. *Ability to score reference signals*. To ascertain that subjects passing the test above do not underscore the reference signals, reference signals where Score < 90 were counted. Subjects showing more than three such scores were rejected. Five subjects failed this test; four of them were already rejected in the previous step. Hence, one additional subject was rejected as a result of this.
3. *Ability to identify and score anchor signals*. To ensure the assessment of anchor signals, two measures were used for each subject: (a) the mean value of Anchor scores and b) the number, n_{EAS}, of scores exceeding a fixed score value. The value used in both measures was Score $= 60$. As a score above this value would represent "good" or "excellent," it is reasonable not to expect such a quality for a majority of Anchor signals. In (b), $n_{\mathrm{EAS}} = 5$ was chosen. When applying any of these conditions, one subject failed, but was already rejected in step 1 above.
4. *Score range used by subjects*. To accommodate for the recommendation on "less critical" and/or "too critical" subjects, the remaining subjects' scores were analysed for each combination of codec and excerpt, yielding 100 combinations. The interquartile range, IQR, was calculated for the scores of each combination. A score outside the range 3IQR from the median was considered an extreme value. The number of extreme values per subject was checked and in no case did any subject show a number exceeding 4 out of the 100 judgements made. This was considered a valid performance. Hence, no subjects were rejected on this criterion only.

In total, eight subjects were rejected as a result of postscreening. Both test sites were equally represented among rejected subjects and the difference in median age between rejected and non-rejected subjects was 0.5 year (31.5 and 31, respectively). The data set now contained 2200 data points (22 subjects * 10 systems * 10 excerpts).

2.2.8 Data analysis

In Test 2, Basic Audio Quality represented by mean values of the dependent variable Score for the different conditions together with the associated 95% confidence intervals were calculated in accordance with recommendations in MUSHRA. These were investigated for significant differences. The systems' compliance with the transparency criteria was also investigated (see Section 3).

2.2.9 Multiple comparisons of systems

In this section, the performance of the different systems across all excerpts except the Electronic (att) is shown (Fig. 9). The Electronic (att) data was excluded from the figure because this excerpt was inserted into the systems at an attenuated level and therefore did not reflect normal broadcasting operating conditions. The results for this excerpt (Appendix, Fig. 13) confirm that it was perceived differently from the other excerpts by exhibiting low scores on all systems due to a high noise level and thus would be unfair to include. For the remaining systems, clear differences can be seen (Fig. 9). An expected observation is that an increased bitrate causes an increase in perceived quality. Across excerpts, the FM 1 system scores equal to DAB+ at 192 kbit/s, whereas FM 2 scores worse than FM 1.

In the Appendix the performance of the different systems for each excerpt is shown together with a brief analysis including possible explanations to system performance. The results show that the different excerpts give rise to different intersystem performance. In summary, the following observations were made:

- In most cases, at higher bit rates no significant differences were observed, that is, the confidence intervals overlap. The MUSHRA method may be less sensitive to the smaller differences at high bit rates.
- In several cases, there was confusion between reference and coded signals.
- Certain signals were more critical, that is, panned speech and transient sounds also occurring in modern music.
- Monotonically increasing scores for increasing bit rate were observed with a few exceptions. This was probably partly caused by the general increase in audio bandwidth and decrease in quantization distortion that follows with higher bit rate.
- FM 1 outperformed DAB+ at low bit rates.
- FM 2 was more sensitive to certain excerpts, probably due to its different clipping algorithm in combination with its narrower bandwidth.
- For seven out of the 10 excerpts the confidence interval of the anchor score overlapped the confidence interval of one or more of the other codecs in the test.

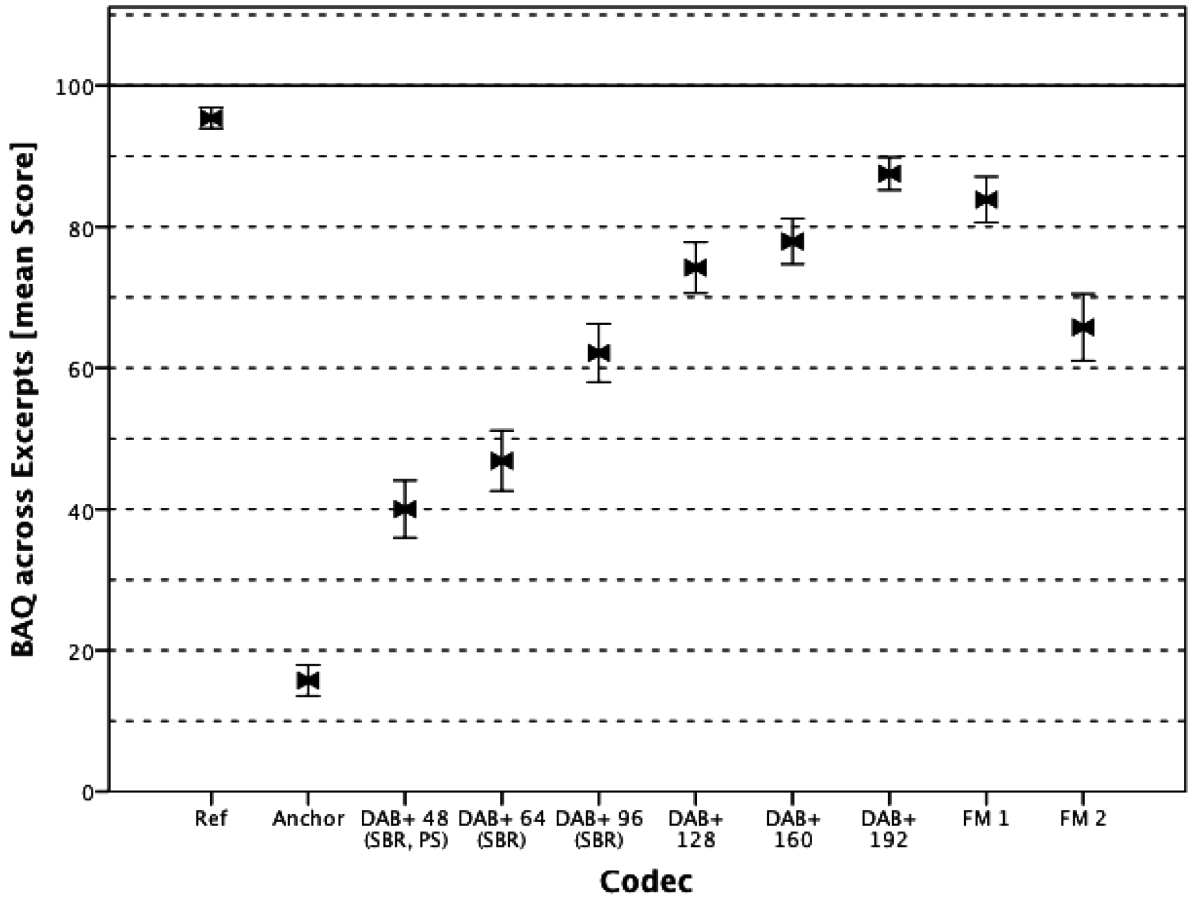


Fig. 9. Mean Basic Audio Quality and 95% confidence intervals for system across all excerpts except Electronic (att).

Comparisons with the audio quality criteria will be discussed in Section 3.

2.3 Interviews

Interviews were conducted (one interview per subject) after the listening tests. For the purpose of this paper, four of the interviews, two from each test site, were randomly selected for analysis. A more extensive analysis is planned for a forthcoming paper. The methodology used had a phenomenological approach for identifying the subject's experience [28]. In this study, this refers to the underlying structure of what degradations the subjects perceived and how the degradations were perceived. An implication of applying a qualitative method is that the results cannot necessarily be applied to the general public, but it gives an insight into understanding the subjects' experience of the tested audio's perceived qualities. Thus, the method can complement the statistical results with information that may be used for coming studies and experimental design, even if it does not strive for completeness or for generalizability.

An interview guide was used to guide the interviewer, and open-ended questions were asked to encourage descriptive statements from the subjects. The questions included the following topics: perceived degradations, what gave low and high scores, the listening training at home, differences between the two tests (BS.1116 and MUSHRA), perception of the scales used in the tests, looping habits, and perceived sound level between the tests. In this paper, a subset of the questions is addressed.

The analysis used meaning condensation as the primary method, where large segments of descriptive transcripts are condensed into short units of meaning, or "the essence" [29],[30]. The units of meaning are then labeled into common themes for all four interviews and then presented as descriptive texts.

The resulting themes from the interviews included information about the perceived degradations, which of them were prominent, whether they were easy/hard to perceive, and what degradations gave low/high scores during the tests. The analysis produced the following descriptive text.

The perceived degradations were changes in the frequency response, comb filter effects, low bit rate artifacts, changes in the stereo image, longer attack time of transients, more noise, and the degradation of the "feeling" of the sound. The most prominent degradations were changes in the stereo image and changes in the frequency response. Degradations caused by noise were easy to perceive, for example, general noise levels, static noise in silent parts of the stimuli. Also changes in the stereo image were prominent. Degradations were perceived when the voice's stereo panorama position was off-centre, as well as when panned transients occurred.

Degradations were hard to perceive in complex stimuli that did not contain any easily discernable reference points. Some subjects focused on the general audio quality,

Table 9. Compliance with quality criterion 1 and 2 for the different systems in Test 2.

	DAB+ (kbit/s)						FM	
Excerpt	48	64	96	128	160	192	1	2
Applause w announcer				×		⊗	⊗	
Classical				×		⊗	⊗	
Electronic				×	×	⊗		
Electronic (att)								
House					⊗	⊗		○
PopKent					⊗	⊗	⊗	⊗
PopRox			⊗			⊗	⊗	
SpeechL (no pan)		⊗	⊗	⊗	⊗	⊗	⊗	
SpeechL (pan)							⊗	
World			×	⊗	⊗	⊗	⊗	⊗
Mean value for systems [Electronic (att) excluded]						×	×	

○ Confidence interval of reference overlaps confidence interval of item (Criterion 1).
× Mean is over 80 (Criterion 2).
⊗ Both criteria are fulfilled.
Blank: None of the criteria are fulfilled.

including listening to the attack of the transients and stereo image to discern the quality of that stimulus. When distributing high scores, the subjects listened for the presence of a sound; one listener associated this with listening to a mix (as if an audio engineer). A sort of weighing strategy was also employed where the subjects listened for the total quality of the degradations to discern whether they were acceptable or whether the stimuli contained other advantages.

Several degradations were perceived as acceptable, for example, a low amount of artifacts induced by the coding algorithm and loss of attack in the transients. One listener even stated that losses of high frequencies are more acceptable compared to coding artifacts, quantizing noise, and low bit depth.

Low scores were given when the following conditions were met:

- a large difference between the stimulus and the reference;
- degradations were audible in combination with how much it affected the result; for example, speech was not scored the lowest because the information in the speech still came through to the listener;
- the bandwidth was extremely limited and when the bandwidth changed with the signal; and
- loss of high frequencies, high-frequency tones, and loss of attack of the transients were perceived.

3 COMPLIANCE WITH AUDIO QUALITY CRITERIA

In this section, the compliance of the systems studied in Test 1 and Test 2 with the previously discussed audio quality criteria 1 through 3 (perceptual transparency, broadcast quality and FM quality) is presented.

3.1 Perceptual transparency and broadcast quality

In Test 1, none of the systems was found to be perceptually transparent (Criterion 1). For the bit rates 96, 128, and 160 kbit/s, none of the tested items fulfilled this criterion. At 192 kbit/s, the items "Applause w announcer" and the two speech items did not reach perceptual transparency. The broadcast quality criterion (Criterion 2) was fulfilled in Test 1 only at 192 kbit/s, but since panned speech at this bit rate did not reach a SDG > -1.0, this system should be used with caution. For Test 2, a similar analysis was made (Table 9).

As is shown by the the table, none of the tested systems were fully perceptually transparent. At 192 kbit/s, all tested items except Electronic (att) and panned speech, however, fulfilled this criterion. As mentioned previously, the Electronic (att) excerpt does not reflect normal operating conditions and should be disregarded, whereas panned speech is commonly used in broadcasting and therefore should be considered an important excerpt.

Table 9 show that both 192 kbit/s and FM 1 in Test 2 fulfilled the broadcast quality criterion. The average score across items for these systems was above 80, but since some individual items score less than 80, these systems should be used with care according to the discussion on broadcast quality in the introduction.

3.2 FM quality

To test specifically the similarity between the systems DAB+ and FM (Criterion 3), paired two-tailed t tests ($df = 21$) were performed on the mean difference $\bar{d}$ between the scores for DAB+ and FM1 for each of the bit rates of DAB+ systems and excerpts except Electronic (att), $d =$ Score(DAB+, [bit rate, excerpt]) $-$ Score(FM 1, [excerpt]). This was repeated for FM 2 replacing FM 1. The effect of multiple comparisons was controlled by applying the false discovery rate (FDR) procedure on p values from all t tests based on $\alpha = 0.05$ [31].

Table 10. *t* tests of differences between DAB+ and FM 1 for excerpts except Electronic (att).

	DAB+ systems at different bit rates compared with FM 1 (kbit/s)					
Excerpt	48	64	96	128	160	192
Applause w announcer	−	−	−	−	−	
Classical	−	−	−			
Electronic	−	−	−			+
House	−	−	−			+
PopKent	−	−	−	−		
PopRox	−	−			−	
SpeechL (no pan)	−					
SpeechL (pan)	−	−	−	−	−	
World	−	−				

Significant differences are denoted by "+" for a higher score ("−" for a lower) of DAB+ systems.

Table 11. *t* tests of differences between DAB+ and FM 2 for excerpts except Electronic (att).

	DAB+ systems at different bit rates compared with FM 2 (kbit/s)					
Excerpt	48	64	96	128	160	192
Applause w announcer	−		+	+	+	+
Classical		+	+	+	+	+
Electronic	−	−	−			+
House	−	−	−			+
PopKent	−	−	−	−		
PopRox	−					+
SpeechL (no pan)		+	+	+	+	+
SpeechL (pan)	+	−	−	−		+
World	−	−	−			

Significant differences are denoted by "+" for a higher score ("−" for a lower) of DAB+ systems.

For FM 1 (Table 10), the results showed that in only two out of the nine cases at 192 kbit/s the DAB+ system was perceived as significantly better, whereas at lower bit rates either no difference or worse performance by the DAB+ was noted. For FM 2 (Table 11), the results were more dispersed: at higher bit rates, DAB+ surpassed FM 2; at lower bit rates, DAB+ fell below; and in a number of cases, no significant differences were found. There were instances where DAB+ was superior to FM 2 down to 64 kbit/s but also inferior up to 128 kbit/s.

3.3 Bit rates required for transparency

To predict the bit rate necessary to attain perceptually transparent quality (i.e., imperceptible difference from the reference), which in Test 1 means that SDG = 0, a model for SDG as a function of bit rate, R_B, in the form of a linear curve, SDG = $bR_B + c$, was fitted onto the data in Test 1. As the excerpts formed the three subsets (see Section 2.1.11) ES1 (SpeechL excerpts), ES2 (House excerpt), and ES3 (Applause and Classical excerpts), these were treated individually. The resulting parameters (b and c), the goodness-of-fit (R^2), as well as the required bit rate for fulfilling SDG = 0 was calculated for each of the subsets (Table 12). The fits of other models were tested, but neither of them showed a superior fit as compared to the fit of the linear model.

Table 12. Parameters of the linear model SDG = $f(R_B)$ and required bit rate for SDG = 0.

Excerpt subset	b	c	R^2	R_B for SDG = 0 [kbit/s]
ES1	0.018	−5.13	0.43	284
ES2	0.019	−3.93	0.35	207
ES3	0.010	−2.30	0.16	222

The results showed that the linear model pointed towards necessary bit rates above 200 kbit/s for any of the subsets to reach SDG = 0. The highest bit rate required was found for subset ES1, which contained the most critical items, where SDG = 0 was reached at R_B = 284 kbit/s. This subset also yielded the best fit ($R^2 = 0.43$) of the model.

4 DISCUSSION

4.1 Audio quality and transparency

Considering the results of the experiments in relation to the audio quality criteria presented in the introduction, several observations can be made.

The preferred subchannel bit rate for the items in Test 2 transmitted over DAB+ should be at least 192 kbit/s for broadcast quality (as single items below Score = 80 were allowed). For perceptual transparency, even higher bit rates would be required due to the lack of compliance of the SpeechL (pan) item. Such bit rates would pose a problem as DAB+ currently only allows for a maximum of 192 kbit/s.

If the results of Test 1 are considered and the definition for broadcast quality is applied, 192 kbit/s would be required, which supports the findings above. Still it has to be noted that the panned speech items did not reach a sufficient level of quality. If the perceptual transparency criterion is applied and the linear model in Section 3.3 is used, this points towards a necessary bit rate above 200 kbit/s, and for the most critical items (panned speech), a bit rate close to 300 kbit/s (Table 12 indicates at least 284 kbit/s) would be needed. It has to be pointed out that the extrapolation by means of a linear function may contain errors that can affect the accuracy of the predicted necessary bitrate. Given the sum of findings, however, there is a need for bit rates to be well above 192 kbit/s, especially for critical items.

In Test 2, if the occurrence of a significant negative BAQ difference between DAB+ and FM for any item was to be applied as a criterion for non-transparency of DAB+ in relation to FM, the following would apply: a DAB+ bit rate of 192 kbit/s would give a result that is a comparable to or better than a modern FM system (FM 1), whereas a bit rate of 160 kbit/s is likely to perform comparable to or better than the average types of FM transmitters used by Swedish Radio (FM 2). Lower bit rates would give rise to significant degradations of the audio quality (Tables 10 and 11).

Certain types of sounds appear quite commonly in regular programming, for example, applause and panned speech

in radio. The quality of such sounds will have a significant weight on the overall perceived quality of a broadcasting channel and will, therefore, be essential to include in quality tests. The occurrence of such signals in a test is also important to consider when averaging scores across items as results from one quite critical item may be masked by less critical ones. Any representation of data from a listening test that just shows the average performance across items may risk missing some of the quality deficiencies of the systems. One example is depicted in Fig. 9 where DAB+ at 128 kbit/s appears to be better than FM 2 on average. However, the analysis of individual items (Table 11) shows that two items have significantly inferior quality. Although the findings in the current study were based on a very limited number of excerpts, they show that sounds that are critical exist as they revealed weaknesses of the systems under test.

A well-known phenomenon is that a codec may have performed well for a vast number of sound excerpts over time, but when exposed to a previously never encountered excerpt, it may produce clearly audible artifacts at the selected bit rate. To reduce such risks when a system's bit rate is to be established and to accommodate for future possible critical items, one solution may be to apply a safety margin in the form of a bit rate that is higher than the one the current listening test indicates as being necessary.

Although the experimental design was aimed to emulate realistic broadcasting conditions, it should be noted that the experiments were performed under more favorable conditions than are normally found in real-life broadcasting. One example is that mobile reception over a long distance from the transmitters may degrade the audio in various ways, both for FM and DAB+. Another example is the common use of cascaded codecs, especially if different bit rates are used at different stages of the cascade. Also, ancillary data may gradually use more bandwidth and thus increase over time, which may in reality lead to a reduction of the available audio bit rate.

It should also be noted that the DAB+ and FM processes were limited to a resolution of 16 bits at the time of the experiments, whereas systems with higher resolution are now available. Another observation about listening tests in general is that some of the reference signals still in use may have been recorded through equipment that now has been surpassed in terms of quality.

As noted previously in this paper, in comparison with a number of other media distribution formats, the bit rates currently used for DAB+ are generally lower. An obvious conclusion is that DAB+ listeners could perceive the system as inferior to other contemporary audio applications.

Altogether the experimental results imply that even higher bit rates may be needed to reach the desired level of quality and transparency in future broadcasting systems.

4.2 Miscellaneous observations

From the interviews, a number of features important for the subjects' assessment of quality were observed. In addition to statements on weighing coding artifacts against reduced audio bandwidth, several details were reported. One particularly interesting detail was related to the appearance of artifacts at stereo panorama positions segregated from the voice's position. More details on these findings will be presented in future publications.

The excerpt Electronic (att) was attenuated so as not to cause clipping in the pre-emphasis circuit. The idea was to compare the excerpt with its unattenuated counterpart (Electronic). However, due to the low signal level and the following make-up gain, the noise level became so loud that the noise itself caught the listeners' attention, which led to significantly lower scores for the attenuated excerpt. This did not correspond to normal broadcasting operating conditions. Additionally, some of the possible artifacts may have been masked by the noise. Consequently, the intended comparison was not possible and had to be abandoned.

In Test 2, the distribution of scores was different between excerpts. In some cases, large quality differences were found; for other cases, the opposite was observed. When the differences are small due to a perceived high quality of the items, the resolution of the MUSHRA method may be insufficient. On the other hand, tests where codecs are compared indirectly by means of a reference, such as the BS.1116, would result in problems where items have large perceptual differences [32].

"FM quality" has been used in a wide range of meanings by the audio community. In this experiment, measures were taken to ensure a high-quality performance within the possible limits. This was accomplished by employment of the recommendations ITU-R BS.412-9 regarding the average multiplex power deviation as well as the peak deviation and EBU R-128 regarding audio levels [22] [26].

As discussed in Section 2.2.1.4, band limited anchor signals are normally used in MUSHRA tests. Typically, the anchor is scored quite low while the other items receive scores that are relatively high in comparison to the anchor due to the anchor's large perceptual difference from the wide-band items [11]. In this experiment, the use of a wide-band frequency spectrum anchor removed that particular effect for several items (Section 2.2.9).

4.3 Future research

The quality issues in broadcasting are numerous and further research is needed in several areas. Some of them are summarized in this section.

Analog FM is likely to be used for many more years in several countries. Further developments in loudness alignment and control of the FM deviation together with new transmitter equipment are expected to improve the FM quality performance even further [22],[26]. This will reinforce FM as a de-facto anchor for audio broadcasting quality that subsequent systems will be compared with, which in its turn puts even more quality pressure on these systems. This fact in combination with the introduction of improved receivers both for FM and DAB+ as well as hybrid receivers for Web radio, DAB+, and FM will raise the question about how these systems relate to one another in terms of quality.

Antenna systems at the transmitter site and also receiver antennas need to be evaluated further. In United Kingdom

and Italy, new national standards for certification of DAB and DAB+ receivers are in the process of being published. These new methods of measuring RF performance and functionality of digital radio receivers, however, do not take into account the perceived audio quality criteria as presented in this paper.

The current study is made with static reception. A long distance to the transmitter may degrade the audio in various ways. For FM, this results in multipath distortion, noise, etc.; for DAB, dropouts and other artifacts may be evident. More efficient error correction methods and ways to increase field strength, for example, by means of transmitter power, may be needed, especially for mobile reception.

Signal loss measurements under realistic reception conditions would give an alternative interesting evaluation of the overall performance of a digital radio system. What would be the listeners' reactions to these types of artifacts?

More tests could clarify the impact of low bit rate systems for contribution and production in cascade with low bit rates used for distribution of digital radio. Test 2 showed that some of the effects of band limited anchor signals were suppressed when using a wideband anchor as defined in Section 2.2.1.4. This is a promising observation and comparison of traditionally used anchor signals with new types of anchor signals should be investigated further.

In summary, the development and refinement of systems for audio broadcasting calls for thorough testing and comparison of the systems in relation to other audio systems available. In such tests, as the findings in this paper show, it will be important to define criteria for minimum acceptable quality and/or transparency. Hence, both test design and criteria for quality decisions are essential components of future research.

5 CONCLUSIONS

To conclude, the systems under test could not be considered as being fully transparent nor could they outperform a realistic optimal FM system on all accounts. The implications of the findings are listed below:

- The subchannel bit rate for DAB+ should not be less than 192 kbit/s for a stereo signal.
- A DAB+ subchannel bit rate of 192 kbit/s would be comparable to or better than the modern FM system.
- A DAB+ subchannel bit rate of 160 kbit/s would be comparable to or better than the average types of FM transmitters used by Swedish Radio.
- Bit rates below these could significantly degrade the quality of certain programme material.
- To accommodate for more critical but still typical items, unless encoding improves, a bit rate close to 300 kbit/s may be necessary for perceptual transparency to be realized.
- When making decisions about broadcasting systems, it will be important to have well-defined criteria for minimum acceptable quality and whether perceptual transparency should be required.

Clearly, the bit rates and encoders in this study would have problems keeping up with the quality of other high-performance audio applications available to the end user. The available bit rates are subject to reduction due to both the number of competing channels that should be fitted as well as the ancillary data included in the bit stream. If broadcasting services advocate and market audio quality as their hallmark, interested parties need to make well-founded decisions on the audio coding infrastructure with respect to contribution, distribution, and emission.

6 ACKNOWLEDGMENTS

The authors wish to thank the participating subjects in the listening tests for their time and effort. Mr. Gerhard Spikofski (IRT) and Mr. Mathias Coinchon (EBU) are thanked for providing data and support from the EBU D/DABA BPN 094 test. The authors are grateful to Mr. Leif Claesson (Claesson-Edwards, Omnia Audio and Linear Acoustic) and Teracom for providing FM equipment. Finally, Dr. Francis Rumsey and Mr. Andrew Mason are acknowledged for their valuable comments.

7 REFERENCES

[1] F. Rumsey (2012). Audio Bit Rates: Downward and Onward. (Feature article). *J. Audio Eng. Soc.*, 60 (9), 729-733.

[2] D. Marston, & A. Mason (2005, October). Cascaded Audio Coding. *EBU Technical Review*.

[3] R. Geiger, R. Yu, J. Herre, S. Rahardja, S. Kim, X. Lin, (2007). ISO/IEC MPEG-4 High-Definition Scalable Advanced Audio Coding. *J. Audio Eng. Soc.*, 55 (1/2), 27–43.

[4] EBU. (2005). *EBU Tech 3309: Evaluations of Cascaded Audio Codecs*. Geneva: European Broadcasting Union.

[5] EBU. (2010). *EBU Tech 3339: EBU Evaluations of Multichannel Audio Codecs, phase 3*. Geneva: European Broadcasting Union.

[6] K. Gross (2011). The Future of High-Performance Media Networking. *AES 44th International Conference: Audio Networking*. New York: Audio Engineering Society.

[7] EBU. (2008). *EBU Tech 3326: Audio contribution over IP: Requirements for Interoperability*. Geneva: European Broadcasting Union.

[8] ISO/IEC. (2004). *International standard 13818-7. Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC)*. Geneva: International Organization for Standardization.

[9] EBU. (1999). *Document BPN 019: Report on the subjective listening tests of multichannel audio codecs*. Geneva: European Broadcasting Union.

[10] EBU. (2009). *Document BPN 094: Subjective Assessment and Objective measurements of DAB+*. Geneva: European Broadcasting Union.

[11] EBU. (2007). *EBU Tech 3324: EBU Evaluations of Multichannel Audio Codecs*. Geneva: European Broadcasting Union.

[12] E. Wyatt (2011, April 21). *A Clash Over the Airwaves*. Retrieved May 27, 2013 from NYTimes.com: http://www.nytimes.com/2011/04/22/business/media/22spectrum.html

[13] ETSI. (2008). *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2 (3GPP TS 36.300 version 8.4.0 Release 8)*. Sophia-Antipolis Cedex: European Telecommunications Standards Institute.

[14] IEEE. (2012). *IEEE Standard for Information technology — Telecommunications and information exchange between systems Local and metropolitan area networks — Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. New York: IEEE.

[15] Wohnort. (2013). *DAB Ensembles Worldwide*. Retrieved June 19, 2013 from http://wohnort.org/DAB/index.html

[16] F. Rumsey (2011). Audio in the age of digital networks. (Feature article). *J. Audio Eng. Soc*., 59 (4), 252.

[17] J. R. Stuart (2004). Coding for High-Resolution Audio Systems. *J. Audio Eng. Soc*., 52 (3), 117–144.

[18] J. Ekeroot, & J. Berg (2008). Audio Software Development - An Audio Quality Perspective. *AES 124th Convention*. New York: Audio Engineering Society.

[19] J. Blauert, & U. Jekosch (2012). A Layer Model of Sound Quality. *J. Audio Eng. Soc*., 60 (1/2), 4–12.

[20] ITU. (1997). *Recommendation ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. Geneva: International Telecommunication Union.

[21] ITU. (2003). *Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems*. Geneva: International Telecommunication Union.

[22] EBU. (2010). *EBU Recommendation R 128: Loudness normalisation and permitted maximum level of audio signals*. Geneva: European Broadcasting Union.

[23] ETSI. (2010). *Digital Audio Broadcasting (DAB); Transport of Advanced Audio Coding (AAC) audio*. Sophia-Antipolis Cedex: European Telecommunications Standards Institute.

[24] ITU. (1990). *Recommendation ITU-R BS.708: Determination of the electro-acoustical properties of studio monitor headphones*. Geneva: International Telecommunication Union.

[25] EBU. (2011). *EBU Tech 3344: Practical guidelines for distribution systems in accordance with EBU R 128*. Geneva: European Broadcasting Union.

[26] ITU. (1998). *Recommendation ITU-R BS.412-9: Planning standards for terrestrial FM sound broadcasting at VHF*. Geneva: International Telecommunication Union.

[27] S. Nielsen, & T. Lundh (2003). Overload in signal conversion. *AES 23rd Conference: Signal Processing in Audio Recording and Reproduction*. New York: Audio Engineering Society.

[28] S. B. Merriam (2009). *Qualitative Research: A Guide to Design and Implementation*. San Francisco: Wiley.

[29] S. Kvale, & S. Brinkmann (2009). *InterViews: Learning the Craft of Qualitative Research Interviewing* (2nd ed.). Thousand Oaks: SAGE.

[30] R. Tesch (1990). *Qualitative Research: analysis types and software tools*. Basingstoke: Falmer Pres.

[31] Y. Benjamini, & Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society. Series B (Methodological)*, 57 (1), 289–300.

[32] G. A. Soulodre, & M. C. Lavoie (1999). Subjective evaluation of large and small impairments in audio codecs. *AES 17th Conference: High-Quality Audio Coding*. New York: Audio Engineering Society.

APPENDIX

RESULTS FROM TEST 2 FOR INDIVIDUAL EXCERPTS

Graphs show Basic Audio Quality (BAQ) as mean value across subjects for individual excerpts. Observations on the results and probable causes are in text referring to graphs.

APPLAUSE WITH ANNOUNCER

The original audio was slightly band-limited, which may be the cause of the similar scores at the higher bit rates. It also had impulses that contained large amounts of high frequencies that were clipped in the FM systems. The difference between FM 1 and FM 2 was probably caused by the difference in method of clipping. The traditional pre-emphasis clipping on the left/right audio in FM 2 was probably the cause of its low score. In FM 1, the pre-emphasis clipping was done on the multiplex signal. (Fig. 10.)

CLASSICAL

The excerpt contained brass instruments and castanets with large amounts of high frequencies and transients clipped in the FM systems. As with the previous excerpt,

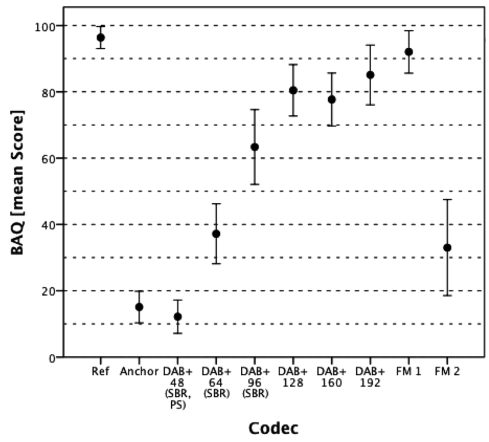


Fig. 10. Means and 95% confidence intervals for Excerpt = Applause with announcer.

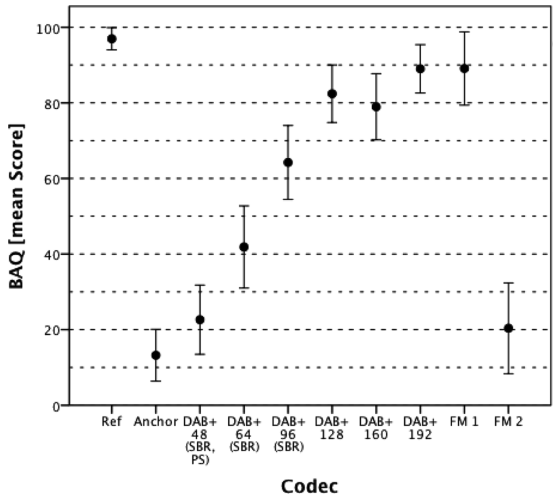


Fig. 11. Means and 95% confidence intervals for Excerpt = Classical.

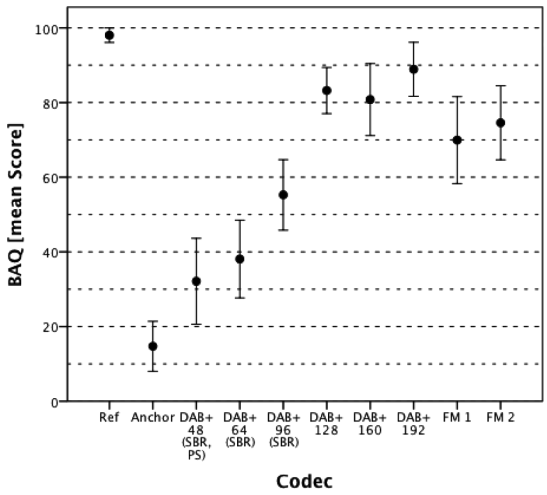


Fig. 12. Means and 95% confidence intervals for Excerpt = Electronic.

the difference between FM 1 and FM 2 was most likely attributable to the clipping method. (Fig. 11.)

ELECTRONIC

The original audio contained quite strong and short transients, which means that they contained a lot of high-frequency components. These components were clipped by the pre-emphasis clippers in the FM systems, a method that probably resulted in the low scores. In contrast to the two previous excerpts, FM 2 received a higher score, which implies that the artifacts were perceived as less severe for this excerpt. (Fig. 12.)

ELECTRONIC (ATTENUATED)

This was the same excerpt as Electronic but attenuated by additionally 22 dB to avoid clipping the FM systems. Unfortunately, the make-up gain applied to this condition resulted in a high noise level that caused low scores for all systems. The main noise source was the FM exciter. (Fig. 13.)

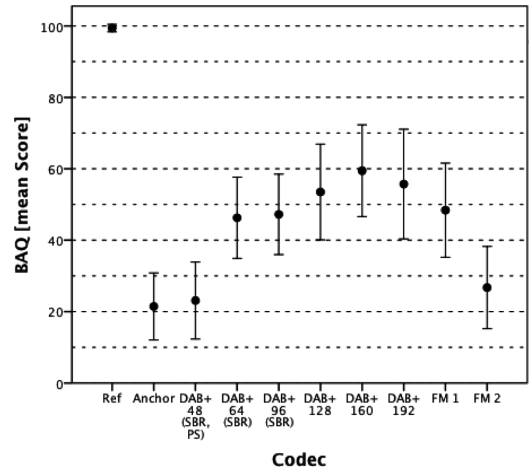


Fig. 13. Means and 95% confidence intervals for Excerpt = Electronic (attenuated).

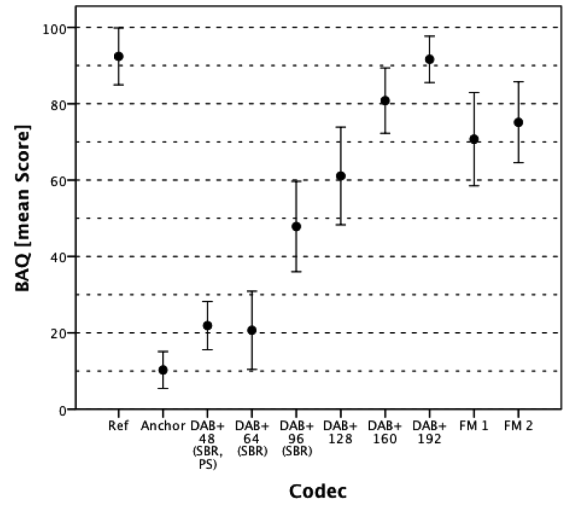


Fig. 14. Means and 95% confidence intervals for Excerpt = House.

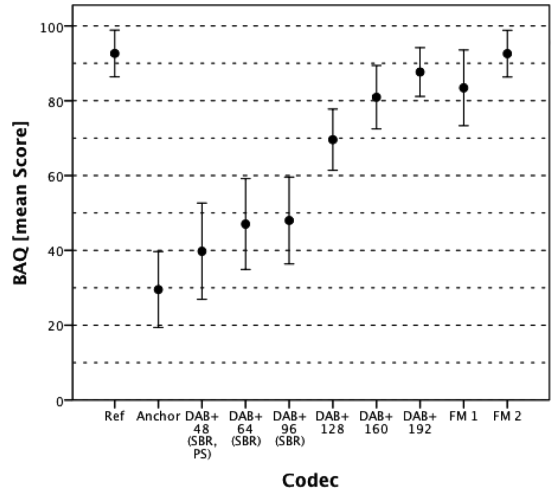


Fig. 15. Means and 95% confidence intervals for Excerpt = PopKent.

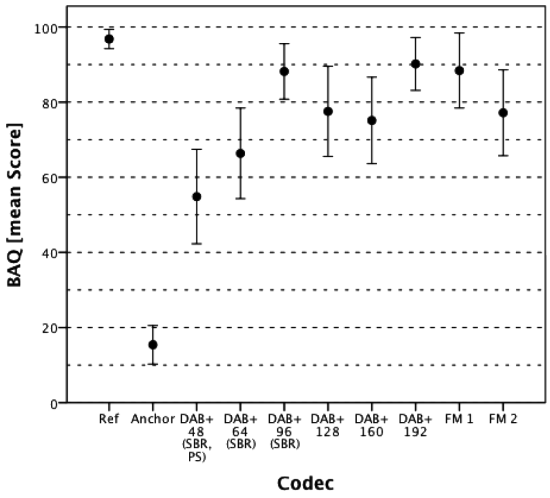


Fig. 16. Means and 95% confidence intervals for Excerpt = PopRox.

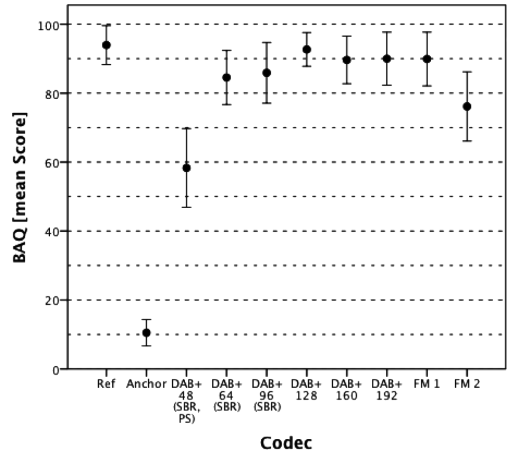


Fig. 17. Means and 95% confidence intervals for Excerpt = SpeechL (no pan).

HOUSE

This excerpt included a bass syntheziser with a complex waveform that contained short spikes that probably caused the low scores at lower DAB+ bit rates. (Fig. 14.)

POPKENT

The excerpt included a panned hi-hat that may have contributed to the low scores at the lower DAB+ bit rates. Except for the hi-hat, the excerpt did not include too much high frequency energy, a condition that proved to be advantageous for FM as well as for AAC. (Fig. 15.)

POPROX

This excerpt had a large amount of high-frequency content. The systems employing SBR yielded a higher audio bandwidth, which was most noticeable at 96 kbit/s. (Fig. 16.)

SPEECHL (NO PAN)

The female speech excerpt was recorded in mono at close distance. Because the voice in this excerpt was centered (equal signals in left and right channel), there was no difference signal ($= L - R$) that had to be coded. Hence, the AAC could allocate all bits to the sum signal. This was probably the reason for the high DAB+ scores, except for 48 kbit/s. (Fig. 17.)

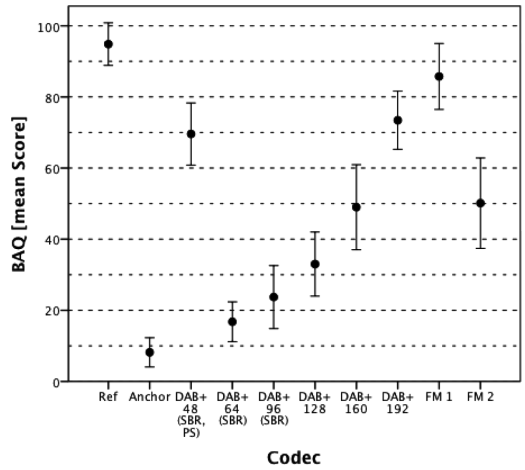


Fig. 18. Means and 95% confidence intervals for Excerpt = SpeechL (pan).

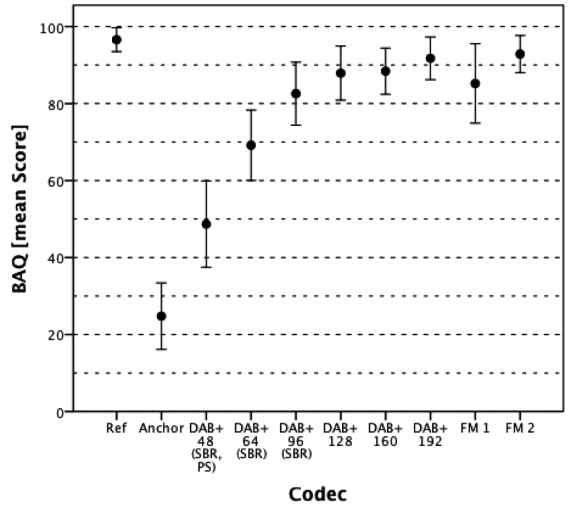


Fig. 19. Means and 95% confidence intervals for Excerpt = World.

SPEECHL (PAN)

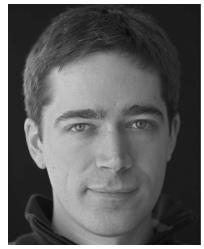
The original audio contained only a single female speaker panned to one side: the interchannel level difference was 6 dB. In this case, using parametric stereo at 48 kbit/s had an advantage. At higher bit rates, parametric stereo was not used, which caused a blur in the stereo image. The traditional pre-emphasis clipping on the left/right audio in FM 2 was probably the cause of the low score. In FM 1, the pre-emphasis clipping was done on the multiplex signal. (Fig. 18.)

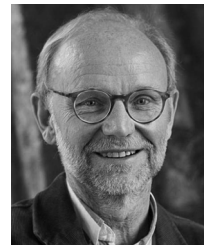
WORLD

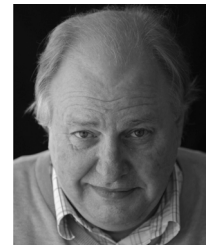
For the World excerpt, the scores followed what may be expected from an increased bit rate. (Fig. 19.)

THE AUTHORS


Jan Berg


Christofer Bustad


Lars Jonsson


Lars Mossberg


Dan Nyberg

Dr. Jan Berg is an associate professor in audio technology at Luleå University of Technology, Sweden, where he carries out research and education on different aspects of audio production with special focus on sound quality evaluation. In his dissertation (2002), Dr. Berg studied evaluation of spatial quality in multichannel systems. Dr. Berg's research activities include collaboration with both academia and industry, for example, in codec testing, loudness, and development of listening tests. Prior to his research work, Dr. Berg worked for more than a decade at Swedish Radio as both recording engineer and maintenance engineer. In 2006, Dr. Berg chaired Sweden's first AES international conference, the AES 28th, on Audio Technology in the Future – Surround and Beyond, taking place in Piteå, Sweden. This was followed in 2010 by the 38th conference on sound quality evaluation. For his work, Dr. Berg received the Board of Governors Award on two separate occasions. Dr. Berg is also active within the AES as a reviewer for the Journal in addition to other duties. He previously held the position as Dean of the Faculty of Arts and Social Sciences at Luleå University of Technology.

●

Christofer Bustad graduated in 2006 with an M.Sc. degree in Engineering Physics from Uppsala University in Sweden with one year on exchange at Queen's University in Ontario, Canada. He did his master's thesis at Swedish Radio Technical Development in MPEG-1/2 Audio Layer II encoding, and joined Swedish Radio Method and Technical Development in 2006. Mr. Bustad has since then been working in the fields of subjective and objective audio quality, audio coding, loudness, FM processing, etc.

●

Lars Jonsson received his M.Sc. degree in Electronic Engineering at the Royal Institute of Technology, KTH, in Stockholm in 1972. Since then he has worked within the Swedish Public Service TV and Radio Broadcasting research and development department. During the last decades, Mr Jonsson's main areas of work have been audio coding and audio quality, digital radio, archiving, audio computer infrastructure, and IP network projects within Swedish Radio. In his international field of work, he has been active in many standard working groups within the Audio Engineering Society and the European Broadcasting Union, EBU. He is vice chair of AES Technical Council Committee on Transmission and Broadcasting. Until recently, he has been the chairman of the EBU Audio Contribution over IP working group, ACIP, and vice chair of EBU Strategic Program on Future Networks and Storage.

●

Lars Mossberg has more than 40 years of experience in broadcasting technologies. After his exam from senior high school, he started at Swedish Radio where he received his two-year training as an audio engineer, a task that he performed for more than 10 years. In 1984, he continued at the company's department for technology development, where he currently works. Mr. Mossberg's main area of professional interest comprises strategic development of audio production and distribution. He has been actively involved in the teams around several listening tests from the early days of peceptual coding to contemporary tests.

●

Dan Nyberg received his Masters Degree in Audio Technology from Luleå University of Technology in 2009. He also received his licentiate degree from Luleå University of Technology in 2013. His licentiate thesis investigates the use of a qualitative data collection methodology in perceptual audio evaluation. He has since then started working with forensic audio at the Swedish National Laboratory of Forensic Science – SKL.