EBU listening tests on Internet audio codecs

G. Stoll
IRT

F. Kozamernik
EBU

The advent of Internet multimedia has stimulated the development of several advanced audio and video compression technologies. Although most of these developments have taken place outside the EBU, many members are using these low bit-rate codecs extensively for their webcasting activities, either for downloading or live streaming. To this end, the EBU Project Group, B/AIM (Audio in Multimedia), was asked to carry out some tests on several low bit-rate audio codecs that are now available on the commercial Internet market.

This article gives the results of the subjective evaluations undertaken by B/AIM in late 1999 and early 2000. These EBU tests are the first international attempt at comparing the different audio compression schemes used on the Internet. In addition, prior to conducting these tests, no internationally-agreed subjective method was available for carrying out evaluations on very low bit-rate, intermediate-quality, codecs. In order to overcome this problem, the group was instrumental in devising a novel test method to evaluate specifically these low-quality audio codecs. The new method is now known as MUSHRA. Both the EBU and ITU-R have now adopted MUSHRA as a standard evaluation method.

1. Introduction

During the last ten years or so, audio coding technology has made enormous progress. Many advanced coding schemes have been developed and successfully used in radio broadcasting, in storage media (e.g. CD, MiniDisc, CD-ROM, DVD) and, particularly, over the Internet. There have been significant advances in terms of the bit-rate reduction achieved, and the quality of the speech and music reproduced has been steadily improv-

ing. Nevertheless, the biggest push in low bit-rate audio coding has taken place quite recently, due to the fast development of the Internet where extremely low bit-rates are required while preserving the subjective quality of the original signal. Digital radio broadcast networks and audio automation systems are now almost completely based on relatively low bit-rate audio coded formats. Within the next few years, the on-line sales and distribution of music may surpass conventional physical distribution channels in terms of market share.

2. Audio codecs market

Following the development of early digital codecs such as NICAM [1] and later ISO/IEC MPEG 1 [2], which are both successfully used in digital broadcasting, there are currently a large variety of different ultra low bit-rate audio codecs, specifically designed for the Internet market. *Table 1* gives a provisional list of the more important codecs. Because of the limited bandwidth available over the Internet, extremely efficient compression techniques for data reduction have been developed.

Current audio-coding standards were developed with relatively simple goals in mind: to achieve the lowest possible data rate while preserving the subjective quality of the original signal. The foreseen applications were digital broadcast emissions (including DAB and DVB), CD-ROM, DVD, etc. Since these channels assume to provide evenly-distributed single errors, error mitigation was limited to simple error detection codes which would allow muting or interpolation of the error-affected frames at the receiver. In the case of the Internet, the error characteristics are “block” in nature and radically lower bit-rates are used, so different design approaches were necessary for optimizing the audio quality at very low bit-rates. Consequently, many new coding schemes were developed specifically for the Internet.

The most advanced audio compression systems spread small portions of the encoded signal – both in time and frequency – and transmit these elements interleaved and spread among many transmission datagrams. Thus the audible effect of a lost or delayed packet can effectively be minimized by interpolating the data between neighbouring packets. In order to make the transmitted stream more robust, some redundancy can be added and the critical elements of the signal can be sent multiple times.

There are additional requirements for advanced compression codecs:

- ⇨ cut-and-paste editing of the encoded format directly, without audible impairments, must be possible;
- ⇨ it should be possible to transmit the same file at different bit-rates, in order to adapt dynamically to network throughput and congestion.

The latter feature is extremely important as it enables optimal sharing of the bit-rate between audio and video, and allows storage of a single file in the content database for a

variety of applications – low bit-rate previews, several different medium bit-rates for streaming, and a high bit-rate version for download or purchase.

As more and more content becomes organized into on-line databases, there is increasing demand for efficient ways to search and categorize this content, and to package it for consumption. It is necessary to index and create metadata using audio analysis tools which classify many parameters of an audio signal. These tools can detect pitch, dynamics, key signature, whether or not the signal contains voice or a musical instrument, how similar the voice is to another voice, etc. Coded formats must support efficient classification. With the adoption of Apple's *QuickTime* as the basis of the ISO MPEG-4 file and

Table 1
Most popular streaming audio and/or video systems *(status: June 1999)*.

	Product Name	Company	Audio/Video	Platform
1	**Advanced Audio Coding (AAC) – MPEG-4**		A	
2	**Audioactive**	Telos	A	Win, Mac
3	**AudioSoft**	Eurodat	A	Win, Mac
4	**Destiny Internet**	Destiny Software	A	Win
5	**Command Engine (DICE)**		I	
6	**I-Media**	Q-Design	A	Win
7	**Intel Streaming Media**	Intel	A/V	Win
8	**Internet Wave**	Vocaltec	A	Win
9	**InterVU**	InterVU	A/V	Win, Mac
10	**MP3**		A	Win, Mac
11	**Netscape Media**	Netscape	A/V	Win, Mac, Unix
12	**QuickTime**	Apple	A/V	Win, Mac
13	**RealAudio**	Progressive Networks	A/V	Win, Mac, Unix
14	**ShockWave**	Macromedia	A/V	Win, Mac
15	**Stream Works**	Xing Technologies	A/V	Win, Mac, Unix
16	**TrueSpeech**	DSP Group	A	Win
17	**ToolVox**	VoxWare	A	Win, Mac, Unix
18	**VDOLive**	VDOnet	A/V	Win, Mac
19	**Vosaic**	Univ. of Illinois	A/V	Win, Mac, Unix
20	**Win Media-Player**	Microsoft	A/V	Win

streaming format, there is a strong common standard architecture defined for the next generation of multimedia systems.

The advent of such a large number of audio codecs has brought a radically new approach to standardization. Standards have become less important, since decoders (which are normally simple and do not require a lot of processing power) are downloadable (possibly in the form of a Java applet) to the client machine along with the content.

In the Internet environment there is no longer a need for a single coding system as is the case in conventional broadcasting. Indeed, RealAudio is no longer the only, and not even the main, audio technology used over the Internet.

From the user point of view, it is irrelevant which audio codec is being used – as long as the technical and commercial performance is comparable. Service providers decide which coding scheme to use. One of the advantages of this "deregulated" approach is that decoders can be regularly updated as the technology advances. The user can have the latest version of the decoder all the time. Audio players can be stored in a flash memory and not on a hard disk.

Browsers or operating systems are usually shipped with a few audio plug-ins. New plug-ins can be downloaded easily. The user is no longer restricted to the use of plug-ins that came with the browser but is free to install any new decoder as appropriate.

The business model of audio streaming is likely to change due to the advent of multicasting. Today, ISPs charge per audio stream. In multicasting situations, however, a single stream will be delivered to several users. The user will then be charged according to the occupancy of the servers used. Due to the huge competition in the audio decoder market, audio streamers will be increasingly available for free.

3. Audio quality assessments

One of the principal characteristics of the current Internet audio codecs is that they experience a large variation in terms of the audio quality achieved for different bit-rates and different audio signals. In addition, they vary in terms of cost, the computation power required (real time), complexity of handling, reliability of the server, the service quality (ruggedness against errors), scalability and marketplace penetration.

The main reason for this is that there is no standard. Even in the MPEG family of standards, the implementation of audio encoders is not standardized, allowing for a large variety of possible implementations in the marketplace. Since the encoder is not standardized, some improvements are possible while keeping the user's decoder terminal unchanged.

Analogue sound systems are measured in terms of the signal-to-noise ratio (S/N) and bandwidth, and they exhibit some harmonic distortions and wide-band noise. Typical artefacts of digital Internet audio codecs are not "harmonic"; they are usually less pleasant for the listener and are often more noticeable and disturbing.

In order to assess the quality of an audio signal under controlled and repeatable conditions, subjective listening tests using a number of qualified listeners and a selection of audio sequences are still recognized as being the most reliable way of quality assessment. ITU-R Recommendation BS.1116-1 [3] is used for the evaluation of high-quality digital audio codecs, exhibiting small impairments of the signal. On the Internet [1] however, medium or even low-quality codecs should be acceptable and are unavoidable. Thus, compromises in the audio quality are necessary. The test method defined in BS.1116-1 is not suitable for assessing such lower audio qualities; it is generally too sensitive, leading to a grouping of results at the bottom of the scale.

This is the main reason that EBU Project Group B/AIM proposed a new test method, termed MUSHRA "**MU**lti **S**timulus test with **H**idden **R**eference and **A**nchors" [4] [2]. This method has been designed to give a reliable and repeatable measure of the audio quality of intermediate-quality signals. The method is in the process of being standardized by the ITU-R [5].

4. The EBU MUSHRA method

Regardless of the method used, the conducting of subjective evaluation tests is generally a highly complex time-consuming and costly process which requires very careful preparation and carrying out, followed by statistical processing of the results [3]. Each of these three phases is briefly described below and is contrasted with ITU-R Recommendation BS.1116-1.

1. Other applications that may require low bit-rate codecs – due to low available bandwidths – and which support intermediate audio quality are digital AM (that is DRM - Digital Radio Mondiale), digital satellite broadcasting, commentary circuits in radio and TV, audio-on-demand services and audio-on-dial-up lines.

2. This inelegant name was agreed by the majority of B/AIM members in spite of some reservations concerning the aesthetic appeal of the acronym. However, taking into account the large impairments and poor audio quality encountered, and the need to endure unpleasant and repetitive listening to the numerous test items, this name does not seem so inadequate.

3. While several such methods have recently been developed (e.g. the new ITU-R PEAQ Standard which has been successfully verified at high audio-quality levels), they are not yet mature and reliable enough to be used in large-scale evaluation tests which feature low and intermediate quality audio, such as the tests described in this article.

4.1. How MUSHRA works

Whereas BS.1116-1 uses a "double-blind triple-stimulus with hidden reference" test method, MUSHRA is a "double-blind multi-stimulus" test method with hidden reference and hidden anchors.

The MUSHRA approach is felt to be more appropriate for evaluating medium and large impairments.

MUSHRA also has the advantage that it provides an absolute measure of the audio quality of a codec which can be compared directly with the reference, i.e. the original audio signal as well as the anchors. Such an absolute measure is necessary in order to be able to compare the results with any other similar tests. If the reference is narrow-band (say 7 kHz), then the codecs under test tend to be rated higher, and this may sometimes lead to very misleading results (e.g. the NADIB test results).

In a test involving small impairments, assessors are asked to detect and assess any perceptible annoyance of artefacts which may be present in the signal. A hidden reference signal helps the assessor to detect these artefacts. On the other hand, in a test with relatively large impairments, the assessor should normally have no difficulty in detecting the artefacts and, therefore, a hidden reference is not necessary. The difficulty however arises when the assessor must grade the relative annoyances of the various artefacts. The assessors are asked to judge their degree of "preference" for one type of artefact versus some other type of artefact.

As MUSHRA is intended for evaluating medium and large impairments, the use of a high-quality reference (as used in BS.1116-1) is to be questioned. The perceptual distance between the reference and the test items is expected to be relatively large. On the other hand, the perceptual distances between the test items belonging to different systems may be quite small. Thus, if each system is only compared with the reference, the differences between any two systems may be too small to discriminate between them. Consequently, MUSHRA uses not only a high-quality reference but also a direct paired comparison between different systems. The assessor can switch at will between the reference signal and any of the systems under test. By way of comparison, in BS.1116-1 the assessor is asked to assess the impairments on "B" compared to a known reference "A" and then to assess "C" compared to "A", where B and C are randomly assigned to a hidden reference and the object under test.

Because the assessors can directly compare the impaired signals, they can relatively easily detect differences between the impaired signals and can then grade them accordingly. This feature permits a high degree of resolution in the grades given to the systems. It is important to note, however, that assessors will derive their grade for a given system by comparing that system to the reference signal, as well as to the other signals in each trial.

In the EBU tests, a computer-controlled replay system was used, although other mechanisms using multiple CD or tape machines can also be used. In a given session, the assessor is presented with a sequence of trials. In each trial, the assessor is presented

with the reference version as well as all versions of the test signal processed by the systems under test. For example, if a test contains seven audio systems, then the assessor is allowed to switch instantly among at least ten signals (one "known" reference + seven impaired signals + one "hidden" reference + at least one "hidden" anchor). Depending on the test, more than one anchor might be used.

During an ITU-R Rec. BS.1116-1 test, assessors tend to approach a given trial by starting with a detection process, followed by a grading process. In MUSHRA, assessors tend to begin a session with a rough estimation of the quality. This is followed by a sorting or ranking process and finally the assessor performs the grading process. Since the ranking is done in a direct fashion, the results are likely to be more consistent and reliable than for the BS.1116-1 method.

4.2. Grading process

The grading scale used in the MUSHRA process is different from the one used in BS.1116-1 which uses the five-grade impairment scale given in ITU-R Recommendation BS.562 [6] In MUSHRA, the assessors are required to score the stimuli according to the five-interval Continuous Quality Scale (CQS) [4]. The CQS consists of identical graphical scales (typically 10 cm long or more, with an internal numerical representation in the range of 0 to 100) which are divided into five equal intervals with the following descriptors from top to bottom:

- ⇨ Excellent
- ⇨ Good
- ⇨ Fair
- ⇨ Poor
- ⇨ Bad

The listeners record their assessments of the quality in a suitable form; for example, with the use of sliders on an electronic display (see *Fig. 1*), or by using a pen and paper scale.

4.3. Reference signals

MUSHRA uses an unprocessed original programme material of *full bandwidth* as the reference signal. In addition, at least one additional signal (*anchor*) – being a low-pass filtered version of the unprocessed signal – should be used. The bandwidth of this additional signal should be 3.5 kHz. Depending on the context of the test, additional anchors can be used optionally. Other types of anchors, showing similar types of impairments as the systems under test, can also be used. For example, these types of impairments can include any of the following possibilities:

4. This scale is also used for the evaluation of picture quality (ITU-R Recommendation BT.500-8 [7]).

- ⇨ bandwidth limitation of 7.0 kHz or 10 kHz;
- ⇨ reduced stereo image;
- ⇨ additional noise;
- ⇨ drop-outs;
- ⇨ packet losses.

In the EBU tests, two *anchor* sequences, i.e. low-pass filtered (3.5 and 7 kHz) versions of the unprocessed signals, were used. In BS.1116-1, the known reference is always available as stimulus "A": the hidden reference and the object are simultaneously available but are randomly assigned to "B" and "C".

4.4. User interface

Compared to ITU-R Rec. BS.1116-1, the MUSHRA method has the advantage of displaying all stimuli for one test item at a given bit-rate at the same time (see *Fig. 1*). The assessors are therefore able to carry out any comparison between them directly. The time consumption for the test is significantly lower than for BS.1116 tests.

Figure 1
User interface for MUSHRA tests.

Fig. 1 shows the user-interface which was used for each session. The buttons represent the reference (which is specially displayed on the top left) and all the codecs under test, including the hidden reference and both processed references, i.e. the two anchors. Under each button, with the exception of the button for the reference, a slider is used to grade the quality of the test item according to the continuous quality scale used. For each of the test items, the signals under test are randomly assigned. In addition, the test items are randomized for each subject within a session. To avoid sequential effects, each assessor runs the five sessions in randomized order.

4.5. Selection of assessors

As in BS.1116-1, listening assessors (i.e. evaluators) should have certain experience in listening critically to the sound sequences. Although the impairments caused by the

Internet audio codecs are generally quite high and therefore relatively easy to detect, experience shows that **experienced listeners give more reliable results, and more quickly than non-experienced listeners. However, non-experienced listeners generally become sensitive enough to the various types of artefacts after frequent exposure. There are methods of pre- and post-screening to eliminate assessors that are not able to discriminate between different artefacts with sufficient accuracy.**

4.6. Training phase

In order to get reliable results, it is mandatory to train the assessors at special training sessions in advance of the test. This training has been found to be important for obtaining reliable results. The training should at least expose the assessor to the full range and nature of the impairments and all the test signals that will be experienced during the test. This may be achieved using several methods: a simple tape replay system or an interactive computer-controlled system.

4.7. Test material

The choice of test material is crucial to the success of the tests and is far from being a simple matter. The MUSHRA method uses a selection of ordinary, unprocessed, broad-

Abbreviations

AAC	(MPEG-2/4) Advanced Audio Coding
AIFF	(Apple) Audio Interchange File Format
ASF	(Microsoft) Advanced Streaming Format
CFI	Confidence interval
CQS	Continuous quality scale
DR	*Danmarks Radio* (Denmark)
DVB	Digital Video Broadcasting
DVD	Digital versatile disc
FhG-IIS	*Fraunhofer Gesellschaft – Institut für Integrierte Schaltungen*
IEC	International Electrotechnical Commission
IRT	*Institut für Rundfunktechnik GmbH* (German broadcast engineering research centre)
ISDN	Integrated services digital network
ISO	International Organization for Standardization
ITU-R	International Telecommunication Union, Radiocommunication Sector
MPEG	Moving Picture Experts Group
MUSHRA	(EBU) MUlti Stimulus test with Hidden Reference and Anchors
NICAM	Near-instantaneous companding and multiplexing
NOS	*Nederlandse Omroep Stichting* (Holland)
NRK	*Norsk rikskringkasting* (Norway)
SR	*Sveriges Television Ab* (Sweden)

cast programme sequences – consisting of pure speech, a mixture of speech, music and background noise, and music only. In contrast, BS.1116-1 uses very critical test sequences specifically chosen to "stress" or even "break" the codec tested and to reveal some audible artefacts. The length of the sequences should typically not exceed 20 s to avoid fatiguing the listeners and also to reduce the total duration of the listening tests.

In order to reveal the differences among the systems under test, the material should be sufficiently critical for each system to be tested. Searching for suitable material is often time consuming; however, unless truly critical material is found for each system, tests may fail to reveal differences among systems and may be inconclusive. On the other hand, too-critical signals (e.g. synthetic, rather than "natural" broadcast programmes) which are deliberately designed to break a specific system should not be used. Care should be taken that the artistic or intellectual content of a programme sequence should be neither so attractive nor so disagreeable or wearisome that the assessors are distracted from focusing on the detection of impairments. The choice should reflect the expected likelihood of occurrence of each type of programme material in actual broadcasts [5].

For the purpose of preparing subjective comparison test tapes, the loudness of each excerpt needs to be adjusted subjectively by the group of skilled assessors – the so-called "experts panel" – prior to recording it on the test media. This will allow subsequent use of the test media at a fixed gain setting for all the programme items within a test trial.

For all test sequences, the group of skilled assessors shall convene and come to a consensus on the relative sound levels of the individual test excerpts. In addition, the experts should come to a consensus on the absolute reproduced sound pressure level for the sequence as a whole, relative to the alignment level. A tone burst [6] at alignment signal level may be included at the head of each recording to enable its output alignment level to be adjusted to the input alignment level required by the reproduction channel [8]. The tone burst is only for alignment purposes: it should not be replayed during the test. The sound-programme signal should be controlled so that the amplitudes of the peaks only rarely exceed the peak amplitude of the permitted maximum signal defined in ITU-R Recommendation ITU BS.645 [9] (a sine wave 9 dB above the alignment level).

The number of test items to be included in a test can vary but it should not be too large, otherwise tests would simply be too long. A reasonable number seems to be around 1.5 times the number of systems under test, with a minimum of 5 items per system. Audio sequences should typically be 10 s to 20 s long. All systems should be tested with the same selection of test items.

The performance of a multichannel system, under the conditions of two-channel playback, shall be tested using a reference down-mix. Although the use of a fixed down-mix may be considered to be restricting in some circumstances, it is undoubtedly the most

5. This condition may be fulfilled with some difficulty since the nature of broadcast material may vary from one station to another and may change in time as musical styles and preferences evolve.

6. For example 1 kHz, 300 ms, -18 dBFS

sensible option for use by broadcasters in the long run. The equations for the reference down-mix [10] are given in:

$$L_0 = 1.00L + 0.71C + 0.71L_s$$

$$R_0 = 1.00R + 0.71C + 0.71R_s$$

It goes without saying that the pre-selection of suitable test excerpts for the critical evaluation of the performance of a reference two-channel down-mix should be based on the reproduction of two-channel down-mixed programme material.

4.8. Listening conditions

The listening tests should be conducted under strictly-controlled conditions as specified in Sections 7 and 8 of ITU-R Recommendation BS.1116-1. Either headphones or loudspeakers are allowed. However, the use of both within one test session is not permitted. All assessors must use the same type of transducer.

Individual adjustment of listening level by a assessor is allowed within a session and should be limited within the range of ± 4 dB relative to the reference level defined in BS.1116-1. The balance between the test items in one test should be provided by the selection panel in such a way that the assessors would normally not need to perform individual adjustments for each item. Level adjustments inside one item should not be allowed.

4.9. Statistical analysis

The statistical analysis of the results obtained is perhaps one of the most demanding tasks. Its purpose is to apply some mathematical operations to the raw data obtained, and then present the results in a user-friendly manner.

The assessments for each test condition are converted linearly from measurements of length on the score sheet to normalized scores in the range 0 to 100, where 0 corresponds to the bottom of the scale (bad quality). Then, the absolute scores are calculated as follows.

Calculation of the averages of the normalized scores of all listeners who remain after post-screening will result in the Mean Subjective Scores (MSS).

The first step in the analysis of the results is the calculation of the mean score, $\overline{u}_{jk}$ for each of the presentations:

$$\bar{u}_{jk} = \frac{1}{N}\sum_{i=1}^{N} u_{ijk} \qquad (1)$$

where: $u_{i\,=}$ the score of observer *i* for a given test condition *j* and sequence *k*
N = the number of observers.

Similarly, overall mean scores, $\bar{u}_j$ and $\bar{u}_k$, could be calculated for each test condition and each test sequence.

When presenting the results of a test, all mean scores should have an associated confidence interval which is derived from the standard deviation and size of each sample.

It is proposed to use the 95% confidence interval which is given by:

$$\left[\bar{u}_{jk} - \delta_{jk}, \bar{u}_{jk} + \delta_{jk}\right]$$

where: $$\delta_{jk} = 1.96\frac{S_{jk}}{\sqrt{N}} \qquad (2)$$

The standard deviation for each presentation, S_{jk}, is given by:

$$S_{jk} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jk} - u_{ijk})^2}{(N-1)}} \qquad (3)$$

With a probability of 95%, the absolute value of the difference between the experimental mean score and the "true" mean score (for a very high number of observers) is smaller than the 95% confidence interval, on condition that the distribution of the individual scores meets certain requirements.

Similarly, a standard deviation S_j could be calculated for each test condition. It is noted however that this standard deviation will, in cases where a small number of test sequences are used, be influenced more by differences between the test sequences used than by variations between the assessors participating in the assessment.

Experience has shown that the scores obtained for different test sequences are dependent on the criticality of the test material used. A more complete understanding of system performance can be obtained by presenting results for different test sequences separately, rather than only as aggregated averages across all the test sequences used in the assessment.

For each test parameter, the mean and 95% confidence interval of the statistical distribution of the assessment grades must be given.

5. The EBU tests

The following seven audio codecs were tested:

- ⇨ Microsoft Windows Media 4
- ⇨ MPEG-2 AAC (implementation by FhG-IIS)
- ⇨ MP3 (close to MPEG-1 and MPEG-2 Layer III, implementation by Opticom)
- ⇨ Q-Design Music Codec 2
- ⇨ RealNetworks 5.0
- ⇨ RealNetworks G2
- ⇨ Yamaha SoundVQ

Each of these codecs was tested at five different bit-rates: 16, 20, 32, 48 and 64 kbit/s. The test was divided into five sessions, according to the five different bit-rates used. In each of these sessions (with the exception of Sessions 4 and 5 [7]), all seven codecs were tested.

- ⇨ **Session 1:** codecs at 16 kbit/sec, mono;
- ⇨ **Session 2:** codecs at 20 kbit/sec, stereo;
- ⇨ **Session 3:** codecs at 32 kbit/sec, stereo;
- ⇨ **Session 4:** codecs at 48 kbit/sec, stereo;
- ⇨ **Session 5:** codecs at 64 kbit/sec, stereo.

The test material was partly taken from earlier Internet Radio listening tests, but also comprised completely new material. The test material consisted of critical, but ordinary broadcast material. It contained pure speech, speech together with music or background noise, as well as music only. The length of the sequences was set to a maximum of 17 s, with a typical length of about 12 s.

The audio items shown in *Table 2* were used for the MUSHRA tests.

The bitstreams produced by the encoders under test at the IRT were sent to T-Nova (Berkom) for verification. The bit-rate was checked for each test item by calculating the size of the encoded file according to the length of the sequence.

Then all bitstreams were decoded or replayed for a subjective check of the technical quality of the items. This was done in order to find any errors which were not caused by the encoding-decoding process. By doing this, an additional check of the bit-rate, as shown in the display of the decoder or player, was done.

7. One of the codecs (i.e. RealAudio 5) did not support 48 and 64 kbit/s and could not be tested in Sessions 4 and 5.

Table 2
Audio test items which were selected for the listening tests

	Type of audio content	Audio item	Recorded by	Comments
1	**Classical music**	Mozart: Requiem – beginning of Dies Irae	IRT	New item
2	**Broadcast programme**	Female speech (Dutch) & Music	NOB	Used already by EBU B/IR group
3	**Broadcast programme**	Female speech (Danish)	DR	Used already by EBU B/IR group
4	**Folk music**	Swedish Folk Music	SR	Used in ITU-R tests (ITU-R TG 10/2)
5	**Live broadcast programme**	Ice-hockey commentary	IRT	New item
6	**Jazz music**	Lee Ritenour	GRP-Records	New item
7	**Broadcast programme**	Male speech (Danish)	DR	Used already by EBU B/IR group
8	**Pop music**	Chris Rea – On the beach		New item
9	**Pop music**	Susan Vega – Tom's dinner		Used already in previous MPEG-tests

6. Summary of test results

The EBU listening tests on Internet audio coding schemes confirmed that the new MUSHRA methodology provides small confidence intervals and thus reliable and stable results. The tests also showed that the evaluation results are repeatable and reproducible.

In the following, the main results of each session are described. The main test results are given in *Fig 2*. More detailed results are available in [4].

6.1. Results for 16 kbit/s per mono signal

The results for a bit-rate of 16 kbit/s per mono signal are given in *Fig. 2a*. These results show that the quality provided by all tested codecs at a bit-rate of 16 kbit/s is significantly lower than the subjective quality of the 7 kHz low-pass anchor. Even more, at this

Click here to download larger versions of these charts *(628 KB)*

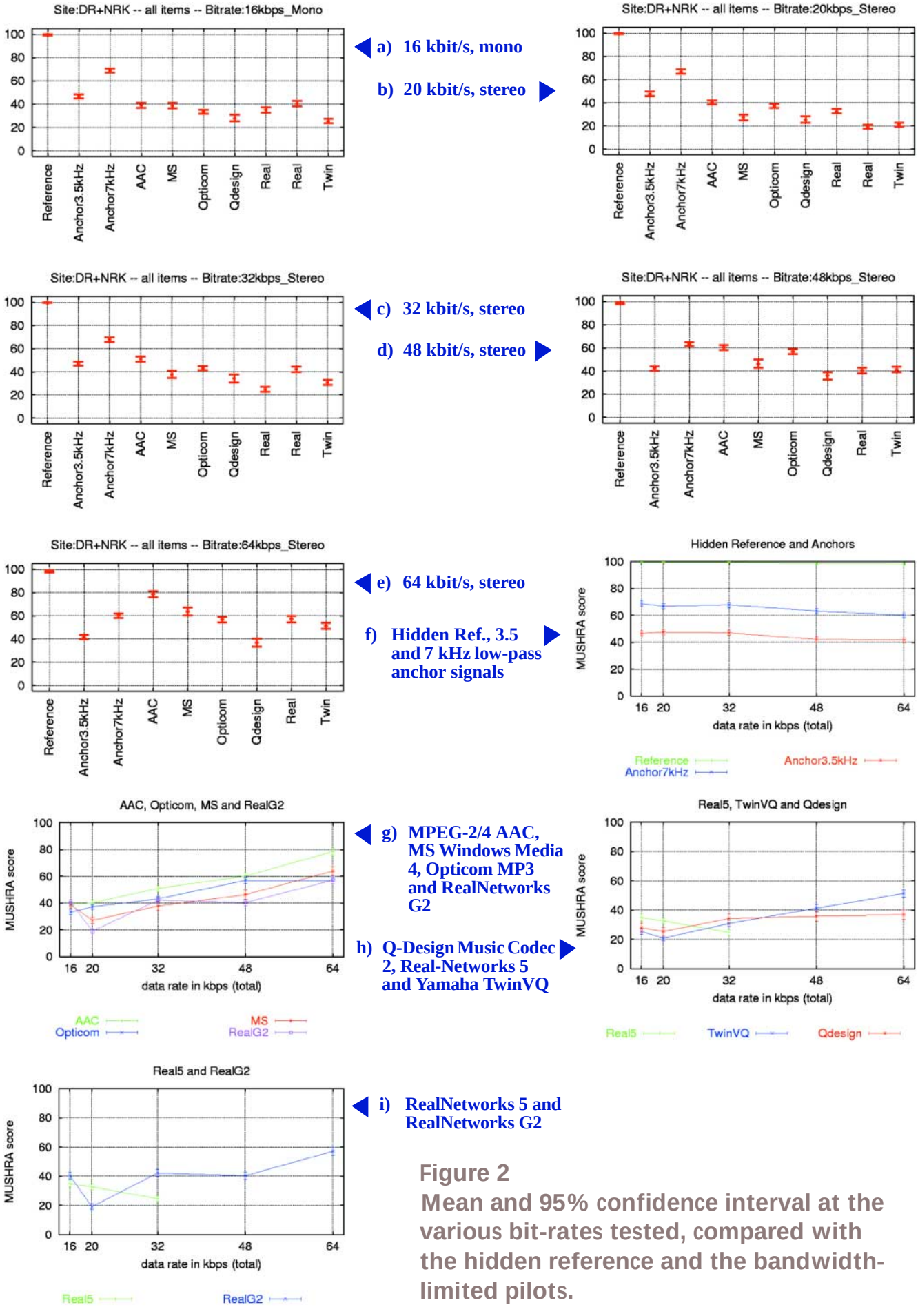

Figure 2
Mean and 95% confidence interval at the various bit-rates tested, compared with the hidden reference and the bandwidth-limited pilots.

bit-rate no codec is better than the 3.5 kHz low-pass anchor. The difference between the different codecs seems to be relatively small, with a grade of about 40 for the best and 25 for the worst.

However, looking at the figures with the detailed results, in particular at those which show the individual test items per codec, it becomes obvious that there are large differences among the different codecs. For example, at 16 kbit/s, the Q-Design Music Codec 2 gives very good quality with all the music-only items. The quality with the folk music item is no different from that of the 7 kHz low-pass anchor, which is in the range of "good quality". The same behaviour can be found for the jazz item. However, this Q-Design codec does not perform so well in cases where music is overlaid by a human voice, or with speech-only items.

6.2. Results for 20 kbit/s per stereo signal

The results for a bit-rate of 20 kbit/s per stereo signal are given in *Fig. 2b*. These results show that the quality provided by all the tested codecs is still significantly lower than the subjective quality of the 7 kHz low-pass anchor. As in the case of 16 kbit/s mono, the quality at 20 kbit/s per stereo signal is also lower than that of the 3.5 kHz low-pass anchor. Comparing the results of Sessions 1 and 2 (i.e. *Figs. 2a* and *2b*), the subjective quality of the 20 kbit/s stereo signal is slightly worse than that of the 16 kbit/s mono signal, for most of the codecs tested. However, in the case of the low-pass filtered anchors, there is no difference between *Figs. 2a* and *2b* (because the only difference between those sessions was that monophonic signals were used in Session 1 and stereophonic in Session 2).

Again, the Q-Design Music Codec 2 showed a very peculiar behaviour. With the two music-only items, it demonstrated good quality. In case of the folk song, the stereo performance was even better than the mono case. However, as soon as human voices were involved in the audio item, the quality of the Q-Design Music Codec 2 dropped significantly.

6.3. Results for 32 kbit/s per stereo signal

The results for a bit-rate of 32 kbit/s per stereo signal are given in *Fig. 2c*. The most obvious result here is that, at this bit-rate, the differences between the various codecs becomes more pronounced. The difference between the best and the worst codec is about 25 points on the 100-point scale, whereas this difference was only about 15 in the case of 16 kbit/s mono. The better codecs are already approaching the subjective quality of the 3.5 kHz low-pass anchor.

6.4. Results for 48 kbit/s per stereo signal

The results for a bit-rate of 48 kbit/s per stereo signal are given in *Fig. 2d*. The MPEG-2/4 AAC and the Opticom MP3 codecs exhibit a "fair" quality level comparable to that of the 7 kHz low-pass filtered anchor. Microsoft Windows Media 4, Q-Design Music Codec 2, RealNetworks G2 and Yamaha TwinVQ are similar to the 3.5 kHz low-pass filtered anchor. It should be pointed out that, for certain audio items (e.g. folk music), the quality of the Windows Media 4 codec was indistinguishable from the hidden reference, whereas the MPEG-2/4 AAC and Opticom MP3 codecs produced a mean value of only 63, i.e. in the range of "good" quality. Considering the results of the Q-Design Music Codec 2, it is interesting to note that the quality at 48 kbit/s did not increase significantly over the quality assessed at 20 kbit/s, for most of the audio items.

6.5. Results for 64 kbit/s per stereo signal

The results for a bit-rate of 64 kbit/s per stereo signal are given in *Fig. 2e*. Several codecs showed very promising results at this bit-rate. In particular, the MPEG-2/4 AAC codec came close to the hidden reference, achieving an overall average of 80 points. It was the only codec in the 64 kbit/s test which was evaluated in the "excellent" range for all the items. Both the MPEG-2/4 AAC codec and the Microsoft Windows Media 4 codec exceeded the quality of the 7 kHz low-pass filtered anchor. The difference between the best and the worst codec was more than 40 points, i.e. the quality differences between the various codecs was greater.

6.6. Results for the hidden anchor and low-pass filtered anchors

As shown in *Fig. 2f*, the Confidence Interval (CFI) for the full-bandwidth reference signal increased at 48 and 64 kbit/s. This was because some of the subjects failed to detect (identify) the hidden reference during the 48 and 64-kbit/s tests. This shows that, even at the relatively low bit-rates considered in these tests, some codecs are capable of offering a quality comparable to the full-bandwidth reference.

In most cases, the CFI of the 7 kHz anchor was evaluated in the range "good" for all the bit-rates tested. The evaluation rating of the 7 kHz anchor however dropped somewhat as the bit-rate was increased, which means that the evaluation of the 7 kHz anchor has some dependency on the bit-rates being evaluated.

The CFI of the 3.5 kHz anchor was evaluated well within the range "fair" at all the bit-rates tested. Again, there was a tendency for the evaluation rating of the 3.5 kHz pilot to drop when the bit-rate was increased. However, the CFI intervals seem to overlap when

comparing the lowest and the highest bit-rates tested, which indicates that the MUSHRA method is an absolute grading system which gives stable and reliable results.

6.7. Mean and 95% confidence interval

Figs. 2g, *2f* and *2i* depict the mean values of the scores and the 95% confidence intervals for the different bit-rates. These charts show that the measurements were very consistent, thus confirming the validity of the MUSHRA method.

7. Main features of the codecs tested

7.1. Microsoft Windows Media 4

This audio system, based on Windows Media Technologies 4.0 and revealed at NAB 99, has two basic codecs that were specifically designed for encoding music and voice content. The encoding speed is rather fast, allowing for real-time encoding on a standard PC, and it can be compared to RealNetworks G2. The multi-threaded architecture increases encoding performance when using more than one processor, i.e. dual-processor systems encode at nearly twice the speed as single-processor systems. MS Media 4 Audio offers a very wide bit-rate range from 5 kbit/s to 128 kbit/s with an 8 kHz to 48 kHz sampling rate, in both mono and stereo. The Media 4 codec is a proprietary system, developed by Microsoft. The version which was tested was an update from August 1999.

For the encoding of voice, Windows Media 4 uses a specially-designed voice codec for compressing the human voice to produce high quality wide-band audio at very low bit-rates. It is based on the ACELP technology and supports bit-rates from 5 kbit/s to 16 kbit/s. This codec was developed by Sipro Lab Telecom.

With Windows Media Technologies version 4.0, content providers can offer as many as five different bit-rates (multi-bit-rate streams) for both on-demand and live streams in a single Advanced Streaming Format (ASF) file. When Windows Media Services and Windows Media Player connect, they automatically determine the available bandwidth. The server then selects and serves the appropriate audio stream. If the available bandwidth changes during a transmission, the server will automatically detect this and switch to a stream with a higher or lower bit-rate.

7.2. MPEG-2, MPEG-4 AAC

AAC forms part of the MPEG-2 and MPEG-4 standards. It uses waveform coding, based on the modified discrete cosine transform (MDCT) of variable length. To prevent

Click here to download larger versions of these charts *(386 KB)*

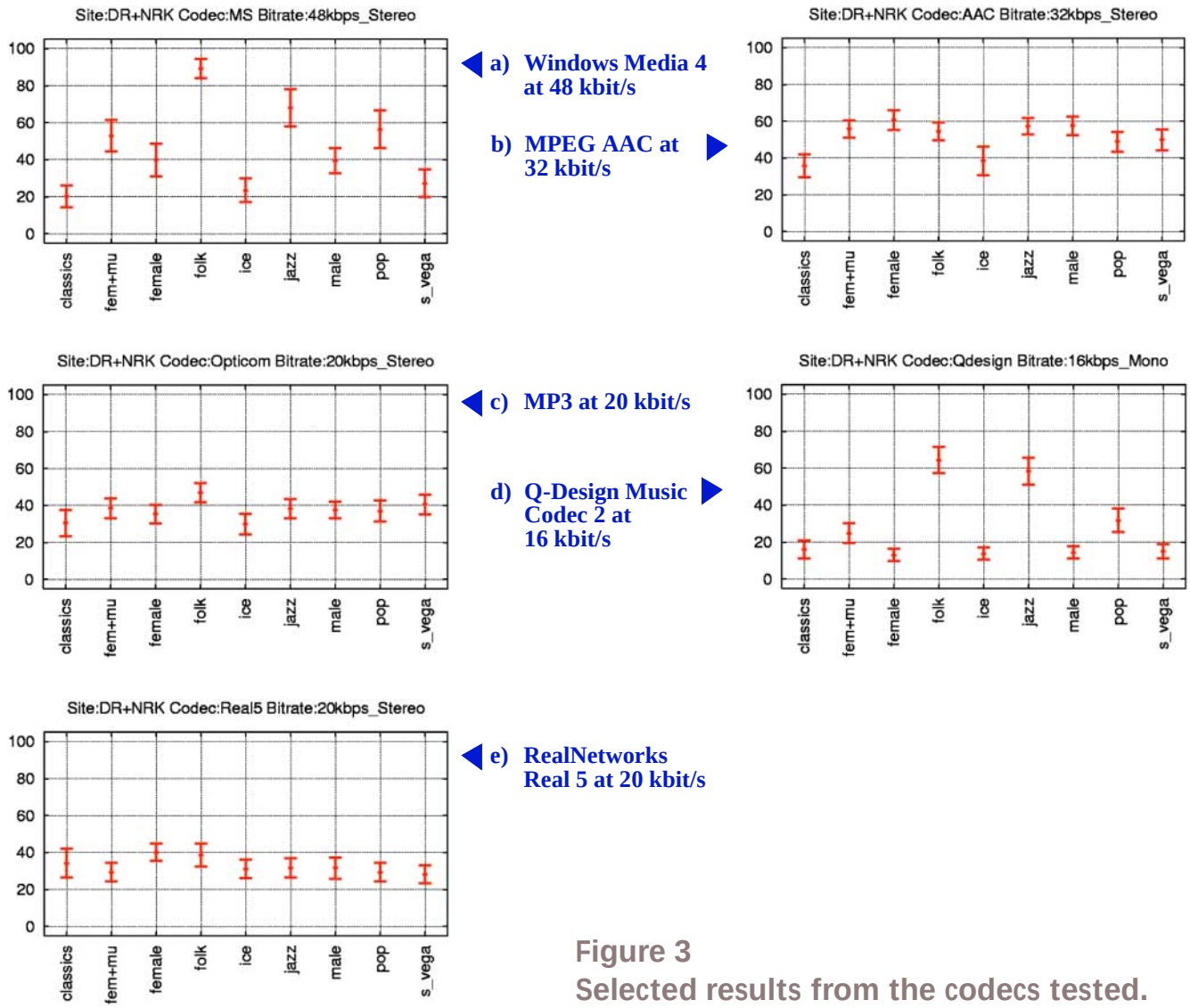

Figure 3
Selected results from the codecs tested.

AAC from becoming a medium for music piracy, AAC is currently only available in secure formats. At present, an Internet application of AAC is only available from Liquid Audio. This specific implementation does not support live streaming nor does it allow replay of AAC-encoded files from normal servers. Currently the system is applicable only to the secure distribution of music over the Internet. In order to prevent music piracy, a specially-certified Liquid Audio server is needed. Other implementations for the use of AAC on the Internet are expected to be available soon. Besides the Internet, AAC will be used in the Japanese HDTV system.

The AAC coder used in this test was the MPEG-2 AAC Main profile encoder according to ISO/IEC 13818-7, implemented by FhG-IIS. AAC was used with four sampling rates between 8 and 32 kHz, depending on the bit-rates in use.

7.3. MPEG-1, MPEG-2 Layer 3 (MP3)

MP3 characterizes a special file format which is mainly used for streaming or downloading of audio files, but also for broadcasting applications (e.g. contributions via ISDN, the

satellite broadcasting system WorldSpace). MP3 is based on the ISO/IEC MPEG Layer 3 standard. There exist several implementations of MP3 encoders and plenty of decoder implementations on the market. The most popular encoders are AudioActive (from Telos Systems), MP3 Producer (from Opticom) and MP3 Live! (from Xing Technologies). All these implementations provide both the standardized sampling rates of ISO/IEC 11172-3 and ISO/IEC 13818-3 and a proprietary extension to very-low sampling rates, named "MPEG-2.5". The MP3 Live! encoder – together with Xing Streamworks MP3 streaming technology, or the AudioActive system using the Microsoft Advanced Streaming Format – are usually taken for live streaming of MP3.

For the EBU tests, Opticom's software encoder and decoder were used. At bit-rates of 48 kbit/s and 64 kbit/s, MP3 was used fully compliant with the MPEG standards whereas at the lower bit-rates, a sampling frequency of 11 kHz (from the MPEG-2.5 extension) was used.

7.4. Q-Design Music Codec 2

This codec runs under the QuickTime 4.0 multimedia platform, which previously was designed only for the downloading of audio and/or video. However, since April 1999 with the first public release of the beta-version of QuickTime 4.0, live-streaming is also supported. The Music Codec 2, is based on a completely new, proprietary, parametric coding system of which details are not available. The public version, which ships without any charge along with the QuickTime 4.0 platform, takes a lot of processing power and thus is very slow. Real-time encoding is more or less impossible with this version. A professional version which automatically adjusts itself to all the necessary refinements involved in audio processing, offers a significantly higher processing speed, allowing for real-time coding on a current standard PC or Mac. A new prototype version was used for the EBU tests, and was not commercially available at the time. The sampling rate was fixed at 44.1 kHz, at all the bit-rates tested [8].

7.5. RealAudio 5.0 and RealNetworks G2

The RealAudio encoder and decoder is a proprietary coding algorithm which supports different coding options with different flavours of the codec.

The RealNetworks G2 audio system is used exclusively for live streaming of audio or the streaming of audio files. However, the creation of WAV or AIFF files is disabled for copy protection reasons. The new G2 system – based on DolbyNet coding technology – provides a big step forward when compared with RealAudio 5.0, thanks to its scalability. To this end, G2 can be used simultaneously on ISDN networks at 64 kbit/s as well as

8. Results below a bit-rate of 32 kbit/s may not be valid for this codec, because a lower sampling frequency might have shown better results.

with a modem of only 14.4 kbit/s capacity. A number of parallel streams, typically up to six, can be created simultaneously within one audio file. The system flexibly allows the quality to be reduced if the available bandwidth reduces (as frequently occurs during Internet rush-hour periods). This facility can be compared to the Intelligent Streaming system used by Windows Media 4.0.

7.6. Yamaha SoundVQ

The Yamaha SoundVQ is a TwinVQ (Transform-domain Weighted Interleave Vector Quantization) coder. It is based on an audio compression technology developed by the NTT Human Interface Laboratories, in which patterns are developed from multiple units of data and compared with standard patterns: compressed code for similar patterns is transmitted. This provides high quality and high compression ratios. The TwinYQ algorithm has been standardized by MPEG-4 Audio. "SoundVQ" is not limited to the distribution of audio data from home pages. It can also be used for voicemail or audio bulletin boards, or for CD-ROMs containing large amounts of audio data. By using the SoundVQ "encoder", anyone can easily create data for distribution. The compression ratio can be selected, allowing the audio data to be compressed from 1/10th to as much as 1/20th of its original size. Since encoded files do not require a special server for distribution, individuals may distribute audio data regardless of their Internet service provider. The "player" is used in conjunction with Internet browsing software, and allows audio to be played back from the user's computer, simply by accessing a homepage.

8. General conclusions

These EBU tests on Internet audio codecs represent a major collaborative achievement among EBU members. They also confirm the well-established EBU role in performing large-scale independent and commercially-neutral evaluations of advanced digital technologies. Following a thorough examination of the test results, the following main conclusions may be drawn:

- ⇨ The AAC codec is the only one in the tests which was evaluated in the range "Excellent" at 64 kbit/s, for all the audio items evaluated.
- ⇨ The Q-Design and RealNetworks 5 codecs produced, over most of the audio items assessed, a grading in the range "Poor" or "Bad", independent of the bit-rate used.
- ⇨ At 16 kbit/s, the Confidence Intervals of the MPEG-2/4 AAC coder are fully or partly within the range of "Fair", except for two items (i.e. Male and Classics). At 64 kbit/s, the Confidence Interval is fully or partly within "Excellent", with the exception of two items (i.e. Ice-hockey and Classics).
- ⇨ MS Windows Media 4 has a quite non-uniform distribution over the different audio items and bit-rates. At 16 kbit/s, the quality varies mainly between the

ranges “Fair” and “Poor”. At 64 kbit/s, depending on the audio item tested, the quality level could be “Excellent”, “Good”, “Fair” or even “Poor”.

⇨ The Opticom codec quality is mainly in the quality range “Poor” at the lowest bit-rate, and mainly “Good” at the highest bit-rate.

⇨ The quality range of the Q-Design Music Codec 2 is very much dependent on the nature of the audio item, and not very much on the chosen bit-rate. The items Folk and Jazz reach a quality level of “Good” even at the lowest bit-rate, but most of the remaining items are placed in the category “Fair” or “Bad” even at the highest bit-rate.

⇨ The RealNetworks 5 codec was tested only at the three lowest bit-rates under test: 16 kbit/s, 20 kbit/s and 32 kbit/s. The quality evaluation of this codec is mainly in the category “Fair” and is independent of bit-rate.

Franc Kozamernik graduated in 1972 from the Faculty of Electrotechnical Engineering, University of Ljubljana, Slovenia. Since 1985 he has been with the European Broadcasting Union (EBU). As a Senior Engineer, he has been involved in a variety of engineering activities, ranging from digital audio broadcasting and audio source coding to the RF aspects of the various audio and video broadcasting system developments. In particular, he contributed to the development and standardization of the DAB and DVB systems.

Currently Mr Kozamernik is the co-ordinator of several EBU research and development Project Groups including B/AIM (Audio in Multimedia) and B/BMW (Broadcasting of Multimedia on the Web). He is also involved in several IST collaborative projects, such as SAMBITS (Advanced Services Market Survey / Deployment Strategies and Requirement / Specification of Integrated Broadcast and Internet Multimedia Services), Hypermedia and S3M.

Franc Kozamernik was instrumental in establishing the EuroDAB Forum in 1994 to promote and roll out DAB, and acted as the Project Director of the WorldDAB Forum until the end of 1999. He represents the EBU in Module A of the WorldDAB Forum. He is also a member of the World Web Consortium (W3C) Advisory Committee.

Gerhard Stoll studied electrical engineering, with the main emphasis on communications theory and psycho-acoustics, at the universities of Stuttgart and Munich. In 1984 he joined the IRT – the research centre of the public broadcasters in Germany, Austria and Switzerland – and became head of the psycho-acoustics group. At the IRT, he was responsible for the development of the MPEG-Audio Layer II standard.

Mt Stoll was/is also a member of different standardizations groups, such as MPEG, Eureka-147, DAB, DVB and the EBU, involved in setting up international standards for broadcasting. For his contributions in the area of low bit-rate audio coding, he received the Prof. Lothar Cremer Award of the German Acoustical Society, and the Fellowship Award of the Audio Engineering Society (AES). As a senior engineer at the IRT, he is now in charge of advanced multimedia broadcasting and information technology services.

⇨ The RealNetworks G2 codec shows at 20 kbit/s a significantly worse quality than at 16 kbit/s mono. At 32 kbit/s it offers a similar quality to 16 kbit/s mono, i.e. it seems that the Real G2 does not gain from any joint stereo coding. Due to the decoded signal's higher frequency response at 48 kbit/s, compared with 32 kbit/s, the quality is even worse than for 32 kbit/s. At 64 kbit/s, the quality is in the range of "Good" and "Fair" for most of the tested signals.

9. Acknowledgements

The authors would like to thank warmly the members of the B/AIM project group who worked hard in conducting the studies, carrying out the subjective tests and putting together the final report which served as the basis for the present article. Particular thanks should go to Messrs. Thomas Sporer (Fraunhofer Institute) for providing the software and user-interface for training and conducting the tests as well as for statistical analysis of the results, Tor Vidar Fosse (NRK) and Michael Harrit (DR) for providing the assessors and for conducting the listening tests, Ulf Wüstenhagen (T-Nova) for verification of test material, and other members of the EBU Project Group B/AIM for their comments and advice.

10. References

[1] ETS 300 163: **Television systems; NICAM 728: Specification for transmission of two-channel digital sound with terrestrial television systems B, G, H, I and L**
http://www.etsi.org/

[2] ISO/IEC 11172-1:1993: **Information technology -- Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s**
http://www.cselt.it/mpeg/standards/mpeg-1/mpeg-1.htm

[3] ITU-R Recommendation BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems
http://www.itu.int/search/index.html

[4] BPN 029: **EBU Report on the Subjective Listening Tests of Some Commercial Internet Audio Codecs**
Contribution of EBU Project Group B/AIM, June 2000.

[5] Preliminary Draft New Recommendation, ITU-R document 10-11Q/TEMP/33:
A method for subjective listening tests of intermediate audio quality - Contribution from the EBU to ITU Working Party 10-11Q
http://www.itu.int/itudoc/itu-r/sg11/docs/wp10-11q/1998-00/contrib/56005.html

[6] ITU-R Recommendation BS.562: **Subjective assessment of sound quality**
http://www.itu.int/plweb-cgi/fastweb?getdoc+view1+itu-doc+12352+1++BS.562

[7] ITU-R Recommendation BT.500: **Methodology for the subjective assessment of the quality of television pictures**
http://www.itu.int/plweb-cgi/fastweb?getdoc+view1+itu-doc+12310+6++BT.500

[8] EBU Recommendation R 68-1992: **Alignment level in digital audio production equipment and in digital audio recorders**
http://www.ebu.ch/tech_texts.html

[9] ITU-R Recommendation BS.645: **Test signals and metering to be used on international sound programme connections**
http://www.itu.int/plweb-cgi/fastweb?getdoc+view1+itu-doc+12361+1++BS.645

[10] ITU-R Recommendation BS.775: **Multichannel stereophonic sound systems with and without accompanying picture**
http://www.itu.int/plweb-cgi/fastweb?getdoc+view1+itu-doc+12373+0++BS.775