RECOMMENDATION ITU-R BS.1534-1

Method for the subjective assessment of intermediate quality level of coding systems

(Question ITU-R 220/10)

(2001-2003)

The ITU Radiocommunication Assembly,

considering

a) that Recommendations ITU-R BS.1116, ITU-R BS.1284, ITU-R BT.500, ITU-R BT.710 and ITU-R BT.811 as well as ITU-T Recommendations P.800, P.810 and P.830, have established methods for assessing subjective quality of audio, video and speech systems;

b) that new kinds of delivery services such as streaming audio on the Internet or solid state players, digital satellite services, digital short and medium wave systems or mobile multimedia applications may operate at intermediate audio quality;

c) that Recommendation ITU-R BS.1116 is intended for the assessment of small impairments and is not suitable for assessing systems with intermediate audio quality;

d) that Recommendation ITU-R BS.1284 gives no absolute scoring for the assessment of intermediate audio quality;

e) that ITU-T Recommendations P.800, P.810 and P.830 are focused on speech signals in a telephone environment and proved to be not sufficient for the evaluation of audio signals in a broadcasting environment;

f) that the use of standardized subjective test methods is important for the exchange, compatibility and correct evaluation of the test data;

g) that new multimedia services may require combined assessment of audio and video quality,

recommends

1 that the testing and evaluation procedures given in Annex 1 of this Recommendation be used for the subjective assessment of intermediate audio quality.

Annex 1

1 Introduction

This Recommendation describes a new method for the subjective assessment of intermediate audio quality. This method mirrors many aspects of Recommendation ITU-R BS.1116 and uses the same grading scale as is used for the evaluation of picture quality (i.e. Recommendation ITU-R BT.500).

The method, called "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)", has been successfully tested. These tests have demonstrated that the MUSHRA method is suitable for evaluation of intermediate audio quality and gives accurate and reliable results, [EBU, 2000a; Soulodre and Lavoie, 1999; EBU, 2000b].

This Recommendation includes the following sections and Appendix:

Section 1: Introduction

Section 2: Scope, test motivation and purpose of new method

Section 3: Experimental design

Section 4: Selection of subjects

Section 5: Test method

Section 6: Attributes

Section 7: Test material

Section 8: Listening conditions

Section 9: Statistical analysis

Section 10: Test report and presentation of results

Appendix 1: Instructions to be given to subjects.

2 Scope, test motivation and purpose of new method

Subjective listening tests are recognized as still being the most reliable way of measuring the quality of audio systems. There are well described and proven methods for assessing audio quality at the top and the bottom quality range.

Recommendation ITU-R BS.1116 – Methods for the subjective assesment of small impairments in audio systems including multichannel sound systems, is used for the evaluation of high quality audio systems having small impairments. However, there are applications where lower quality audio is acceptable or unavoidable. Rapid developments in the use of the Internet for distribution and broadcast of audio material, where the data rate is limited, have led to a compromise in audio quality. Other applications that may contain intermediate audio quality are digital AM (i.e. digital radio mondiale (DRM), digital satellite broadcasting, commentary circuits in radio and TV, audio on demand services and audio on dial-up lines. The test method defined in Recommendation ITU-R BS.1116 is not entirely suitable for evaluating these lower quality audio systems [Soulodre and Lavoie, 1999] because it is poor at discriminating between small differences in quality at the bottom of the scale.

Recommendation ITU-R BS.1284 gives only methods which are dedicated either to the high quality audio range or gives no absolute scoring of audio quality.

Other Recommendations, like ITU-T Recommendations P.800, P.810 or P.830, are focused on subjective assessment of speech signals in a telephone environment. The European Broadcasting Union (EBU) Project Group B/AIM has done experiments with typical audio material as used in a broadcasting environment using these ITU-T methods. None of these methods fulfils the

requirement for an absolute scale, comparison with a reference signal and small confidence intervals with a reasonable number of subjects at the same time. Therefore the evaluation of audio signals in a broadcasting environment cannot be done properly by using one of these methods.

The new test method described in this Recommendation is intended to give a reliable and repeatable measure of systems having audio quality which would normally fall in the lower half of the impairment scale used by Recommendation ITU-R BS.1116 [EBU, 2000a; Soulodre and Lavoie, 1999; EBU, 2000b]. In the MUSHRA test method, a high quality reference signal is used and the systems under test are expected to introduce significant impairments. If the systems under test can improve the subjective quality of a signal then other test methods should be used.

3 Experimental design

Many different kinds of research strategies are used in gathering reliable information in a domain of scientific interest. In the subjective assessment of impairments in audio systems, the most formal experimental methods shall be used. Subjective experiments are characterized firstly by actual control and manipulation of the experimental conditions, and secondly by collection and analysis of statistical data from listeners. Careful experimental design and planning is needed to ensure that uncontrolled factors which can cause ambiguity in test results are minimized. As an example, if the actual sequence of audio items were identical for all the subjects in a listening test, then one could not be sure whether the judgements made by the subjects were due to that sequence rather than to the different levels of impairments that were presented. Accordingly, the test conditions must be arranged in a way that reveals the effects of the independent factors, and only of these factors.

In situations where it can be expected that the potential impairments and other characteristics will be distributed homogeneously throughout the listening test, a true randomization can be applied to the presentation of the test conditions. Where non-homogeneity is expected this must be taken into account in the presentation of the test conditions. For example, where material to be assessed varies in level of difficulty, the order of presentation of stimuli must be distributed randomly, both within and between sessions.

Listening tests need to be designed so that subjects are not overloaded to the point of lessened accuracy of judgement. Except in cases where the relationship between sound and vision is important, it is preferred that the assessment of audio systems is carried out without accompanying pictures. A major consideration is the inclusion of appropriate control conditions. Typically, control conditions include the presentation of unimpaired audio materials, introduced in ways that are unpredictable to the subjects. It is the differences between judgement of these control stimuli and the potentially impaired ones that allows one to conclude that the grades are actual assessments of the impairments.

Some of these considerations will be described later. It should be understood that the topics of experimental design, experimental execution, and statistical analysis are complex, and that not all details can be given in a Recommendation such as this. It is recommended that professionals with expertise in experimental design and statistics should be consulted or brought in at the beginning of the planning for the listening test.

4 Selection of subjects

Data from listening tests assessing small impairments in audio systems, as in Recommendation ITU-R BS.1116, should come from subjects who have experience in detecting these small impairments. The higher the quality reached by the systems to be tested, the more important it is to have experienced listeners.

4.1 Criteria for selecting subjects

Although the MUSHRA test method is not intended to be applied to small impairments, it is still recommended that experienced listeners should be used. These listeners should have experience in listening to sound in a critical way. Such listeners will give a more reliable result more quickly than non-experienced listeners. It is also important to note that most non-experienced listeners tend to become more sensitive to the various types of artefacts after frequent exposure.

There is sometimes a reason for introducing a rejection technique either before (pre-screening) or after (post-screening) the real test. In some cases both types of rejections might be used. Here, rejection is a process where all judgements from a particular subject are omitted.

Any type of rejection technique, not carefully analysed and applied, may lead to a biased result. It is thus extremely important that, whenever elimination of data has been made, the test report clearly describes the criterion applied.

4.1.1 Pre-screening of subjects

The listening panel should be composed of experienced listeners, in other words, people who understand and have been properly trained in the described method of subjective quality evaluation. These listeners should:

– have experience in listening to sound in a critical way;

– have normal hearing (ISO Standard 389 should be used as a guideline).

The training procedure might be used as a tool for pre-screening.

The major argument for introducing a pre-screening technique is to increase the efficiency of the listening test. This must however be balanced against the risk of limiting the relevance of the result too much.

4.1.2 Post-screening of subjects

Post-screening methods can be roughly separated into at least two classes:

– one is based on the ability of the subject to make consistent repeated gradings;

– the other relies on inconsistencies of an individual grading compared with the mean result of all subjects for a given item.

It is recommended to look to the individual spread and to the deviation from the mean grading of all subjects.

The aim of this is to get a fair assessment of the quality of the test items.

If few subjects use either extreme end of the scale (excellent, bad) and the majority are concentrated at another point on the scale, these subjects could be recognized as outliers and might be rejected.

Due to the fact that "intermediate quality" is tested, a subject should be able to identify the coded version very easily and therefore find a grade which is in the range of the majority of the subjects. Subjects with grades at the upper end of the scale are likely to be less critical and subjects who have grades only at the lowest end of the scale are likely to be too critical. By rejecting these extreme subjects a more realistic quality assessment is expected.

The methods are primarily used to eliminate subjects who cannot make the appropriate discriminations. The application of a post-screening method may clarify the tendencies in a test result. However, bearing in mind the variability of subjects' sensitivities to different artefacts, caution should be exercised. By increasing the size of the listening panel, the effects of any individual subject's grades will be reduced and so the need to reject a subject's data is greatly diminished.

4.2 Size of listening panel

The adequate size for a listening panel can be determined if the variance of grades given by different subjects can be estimated and the required resolution of the experiment is known.

Where the conditions of a listening test are tightly controlled on both the technical and behavioural side, experience has shown that data from no more than 20 subjects are often sufficient for drawing appropriate conclusions from the test. If analysis can be carried out as the test proceeds, then no further subjects need to be processed when an adequate level of statistical significance for drawing appropriate conclusions from the test has been reached.

If, for any reason, tight experimental control cannot be achieved, then larger numbers of subjects might be needed to attain the required resolution.

The size of a listening panel is not solely a consideration of the desired resolution. The result from the type of experiment dealt with in this Recommendation is, in principle, only valid for precisely that group of experienced listeners actually involved in the test. Thus, by increasing the size of the listening panel the result can be claimed to hold for a more general group of experienced listeners and may therefore sometimes be considered more convincing. The size of the listening panel may also need to be increased to allow for the probability that subjects vary in their sensitivity to different artefacts.

5 Test method

The MUSHRA test method uses the original unprocessed programme material with full bandwidth as the reference signal (which is also used as a hidden reference) as well as at least one hidden anchor. Additional well-defined anchors can be used as described in Section 5.1.

5.1 Description of test signals

The length of the sequences should typically not exceed 20 s to avoid fatiguing of listeners and to reduce the total duration of the listening test.

The set of processed signals consists of all the signals under test and at least one additional signal (anchor) being a low-pass filtered version of the unprocessed signal. 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 be used. These types of impairments can include:

– bandwidth limitation of 7 kHz or 10 kHz;

– reduced stereo image;

– additional noise;

– drop outs;

– packet losses;

– and others.

NOTE 1 – The bandwidths of the anchors correspond to the Recommendations for control circuits (3.5 kHz), used for supervision and coordination purpose in broadcasting, commentary circuits (7 kHz) and occasional circuits (10 kHz), according to ITU-T Recommendations G.711, G.712, G.722 and J.21, respectively. The characteristic of the 3.5 kHz low-pass filter should be as follows:

f_c = 3.5 kHz

Maximum passband ripple = ±0.1 dB

Minimum attenuation at 4 kHz = 25 dB

Minimum attenuation at 4.5 kHz = 50 dB.

The additional anchors are intended to provide an indication of how the systems under test compare to well-known audio quality levels and should not be used for rescaling results between different tests.

5.2 Training phase

In order to get reliable results, it is mandatory to train the subjects in 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 subject to the full range and nature of impairments and all test signals that will be experienced during the test. This may be achieved using several methods: a simple tape playback system or an interactive computer-controlled system. Instructions are given in Appendix 1.

5.3 Presentation of stimuli

MUSHRA is a double-blind multi-stimulus test method with hidden reference and hidden anchor(s), whereas Recommendation ITU-R BS.1116 uses a “double-blind triple-stimulus with hidden reference” test method. The MUSHRA approach is felt to be more appropriate for evaluating medium and large impairments [Soulodre and Lavoie, 1999].

In a test involving small impairments, the difficult task for the subject is to detect any artefacts which might be present in the signal. In this situation a hidden reference signal is necessary in the test in order to allow the experimenter to evaluate the subject’s ability to successfully detect these artefacts. Conversely, in a test with medium and large impairments, the subject has no difficulty in detecting the artefacts and therefore a hidden reference is not necessary for this purpose. Rather, the difficulty arises when the subject must grade the relative annoyances of the various artefacts. Here the subject must weigh his preference for one type of artefact versus some other type of artefact.

The use of a high quality reference introduces an interesting problem. Since the new methodology is to be used for evaluating medium and large impairments, the perceptual difference from the reference signal to the test items is expected to be relatively large. Conversely, the perceptual differences between the test items belonging to different systems may be quite small. As a result, if a multi-trial test method (such as is used in Recommendation ITU-R BS.1116) is used, it may be very difficult for subjects to accurately discriminate between the various impaired signals. For example, in a direct paired comparison test subjects might agree that System A is better than System B. However, in a situation where each system is only compared with the reference signal (i.e. System A and System B are not directly compared to each other), the differences between the two systems may be lost.

To overcome this difficulty, in the MUSHRA test method, the subject can switch at will between the reference signal and any of the systems under test, typically using a computer-controlled replay system, although other mechanisms using multiple CD or tape machines can be used. The subject is presented with a sequence of trials. In each trial the subject 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 8 audio systems, then the subject is allowed to switch instantly among the 11 signals (1 reference + 8 impaired + 1 hidden reference + 1 hidden anchor).

Because the subject can directly compare the impaired signals, this method provides the benefits of a full paired comparison test in that the subject can more easily detect differences between the impaired signals and 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 subjects 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.

It is recommended that no more than 15 signals (e.g. 12 systems under test, 1 known reference, 1 hidden anchor, and 1 hidden reference) should be included in any trial.

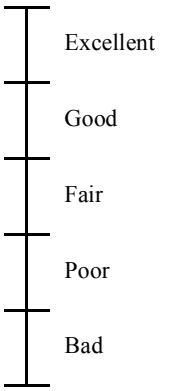
In a Recommendation ITU-R BS.1116 test, subjects tend to approach a given trial by starting with a detection process, followed by a grading process. The experience from conducting tests according to the MUSHRA method shows, that subjects tend to begin a session with a rough estimation of the quality. This is followed by a sorting or ranking process. After that the subject performs the grading process. Since the ranking is done in a direct fashion, the results for intermediate audio quality are likely to be more consistent and reliable than if the Recommendation ITU-R BS.1116 method had been used.

5.4 Grading process

The subjects are required to score the stimuli according to the continuous quality scale (CQS). The CQS consists of identical graphical scales (typically 10 cm long or more) which are divided into five equal intervals with the adjectives as given in Fig. 1 from top to bottom.

This scale is also used for evaluation of picture quality (Recommendation ITU-R BT.500 – Methodology for the subjective assessment of the quality at television pictures).

FIGURE 1



The listener records his/her assessment of the quality in a suitable form, for example, with the use of sliders on an electronic display (see Fig. 2), or using a pen and paper scale. Using a set up similar to that shown in Fig. 2 the subject should be constrained, to be able only to adjust the score assigned to the item he or she is currently listening to. Some guidance about interface design can be found in Appendix 2 to Annex 1. The subject is asked to assess the quality of all stimuli, according to the five-interval CQS.

FIGURE 2

Example of a computer display used for a MUSHRA test

Compared to Recommendation ITU-R BS.1116, the MUSHRA method has the advantage of displaying all stimuli at the same time so that the subject is able to carry out any comparison between them directly. The results become more consistent, leading to smaller confidence intervals. The time taken to perform the test using the MUSHRA method can be significantly less than when using the Recommendation ITU-R BS.1116 method.

5.5 Recording of test sessions

In the event that something anomalous is observed when processing assigned scores, it is very useful to have a record of the events that produced the scores. A relatively simple way of achieving this is to make video and audio recordings of the whole test. In the case where an anomalous grade is found in a set of results, the tape recording can be inspected to try to establish whether the reason was human error or equipment malfunction.

6 Attributes

Listed below are attributes specific to monophonic, stereophonic and multichannel evaluations. It is preferred that the attribute "basic audio quality" be evaluated in each case. Experimenters may choose to define and evaluate other attributes.

Only one attribute should be graded during one trial. When subjects are asked to assess more than one attribute in each trial they can become overburdened or confused, or both, by trying to answer multiple questions about a given stimulus. This might produce unreliable gradings for all the questions.

6.1 Monophonic system

Basic audio quality: This single, global attribute is used to judge any and all detected differences between the reference and the object.

6.2 Stereophonic system

Basic audio quality: This single, global attribute is used to judge any and all detected differences between the reference and the object. The following additional attribute may be of interest:

Stereophonic image quality: This attribute is related to differences between the reference and the object in terms of sound image locations and sensations of depth and reality of the audio event. Although some studies have shown that stereophonic image quality can be impaired, sufficient research has not yet been done to indicate whether a separate rating for stereophonic image quality as distinct from basic audio quality is warranted.

NOTE 1 – Up to 1993, most small impairment subjective evaluation studies of stereophonic systems have used the attribute basic audio quality exclusively. Thus the attribute stereophonic image quality was either implicitly or explicitly included within basic audio quality as a global attribute in those studies.

6.3 Multichannel system

Basic audio quality: This single, global attribute is used to judge any and all detected differences between the reference and the object.

The following additional attributes may be of interest:

Front image quality: This attribute is related to the localization of the frontal sound sources. It includes stereophonic image quality and losses of definition.

Impression of surround quality: This attribute is related to spatial impression, ambience, or special directional surround effects.

7 Test material

Critical material which represents the typical broadcast programme for the desired application shall be used in order to reveal differences among systems under test. Critical material is that which stresses the systems under test. There is no universally suitable programme material that can be used to assess all systems under all conditions. Accordingly, critical programme material must be sought explicitly for each system to be tested in each experiment. The search for suitable material is usually time-consuming; however, unless truly critical material is found for each system, experiments will fail to reveal differences among systems and will be inconclusive.

It must be empirically and statistically shown that any failure to find differences among systems is not due to experimental insensitivity which may be caused by poor choices of audio material, or any other weak aspects of the experiment. Otherwise this “null” finding cannot be accepted as valid.

In the search for critical material, any stimulus that can be considered as potential broadcast material shall be allowed. Synthetic signals deliberately designed to break a specific system should not be included. The artistic or intellectual content of a programme sequence should be neither so attractive nor so disagreeable or wearisome that the subject is distracted from focusing on the detection of impairments. The expected frequency of occurrence of each type of programme material in actual broadcasts should be taken into account. However, it should be understood that the nature of broadcast material might change in time with future changes in musical styles and preferences.

When selecting the programme material, it is important that the attributes which are to be assessed are precisely defined. The responsibility of selecting material shall be delegated to a group of skilled subjects with a basic knowledge of the impairments to be expected. Their starting point shall be based on a very broad range of material. The range can be extended by dedicated recordings.

For the purpose of preparing for the formal subjective test, the loudness of each excerpt needs to be adjusted subjectively by the group of skilled subjects prior to recording it on the test media. This will allow subsequent use of the test media at a fixed gain setting for all programme items within a test trial.

For all test sequences the group of skilled subjects 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 (for example 1 kHz, 300 ms, –18 dBFS) 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, according to EBU Recommendation R 68 (see Recommendation ITU-R BS.1116, § 8.4.1). 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 Recommendation ITU-R BS.645 (a sine wave 9 dB above the alignment level).

The feasible number of excerpts to include in a test varies: it shall be equal for each system under test. A reasonable estimate is 1.5 times the number of systems under test, subject to a minimum value of 5 excerpts. Audio excerpts will be typically 10 s to 20 s long. Due to the complexity of the task, the systems under test should be available. A successful selection can only be achieved if an appropriate time schedule is defined.

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 sensible option for use by broadcasters in the long run. The equations for the reference down-mix (see Recommendation ITU-R BS.775) are:

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

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

The pre-selection of suitable test excerpts for the critical evaluation of the performance of reference two-channel down-mix should be based on the reproduction of two-channel down-mixed programme material.

8 Listening conditions

Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems are defined in Recommendation ITU-R BS.1116. For evaluating audio systems having intermediate quality the listening conditions outlined in Sections 7 and 8 of Recommendation ITU-R BS.1116 shall be used.

Either headphones or loudspeakers may be used in the test. The use of both within one test session is not permitted: all subjects must use the same type of transducer.

For a measuring signal with an r.m.s. voltage equal to the “alignment signal level” (0 dBu0s according to Recommendation ITU-R BS.645; –18 dB below the clipping level of a digital tape recording, according to EBU Recommendation R 68) fed in turn to the input of each reproduction channel (i.e. a power amplifier and its associated loudspeaker), the gain of the amplifier shall be adjusted to give the reference sound pressure level (IEC/A-weighted, slow):

$$L_{ref} = 85 - 10 \log n \pm 0.25 \quad \text{dBA}$$

where n is the number of reproduction channels in the total set-up.

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

Level adjustments inside one item should be not allowed.

9 Statistical analysis

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.

The calculation of the averages of normalized scores of all listeners remaining after post-screening will result in the mean subjective scores.

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

$$\overline{u}_{jk} = \frac{1}{N}\sum_{i=1}^{N} u_{ijk} \tag{1}$$

where:

u_i: score of observer i for a given test condition j and audio sequence k

N: number of observers.

Similarly, overall mean scores, $\overline{u}_j$ and $\overline{u}_k$, 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:

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

where:

$$\delta_{jk} = t_{0.05}\frac{S_{jk}}{\sqrt{N}} \tag{2}$$

and $t_{0.05}$ is the t value for a significance level of 95%.

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

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

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.

10 Test report and presentation of results

10.1 General

The presentation of the results should be made in a user friendly way such that any reader, either a naïve one or an expert, is able to get the relevant information. Initially any reader wants to see the overall experimental outcome, preferably in a graphical form. Such a presentation may be supported by more detailed quantitative information, although full detailed numerical analyses should be in appendices.

10.2 Contents of the test report

The test report should convey, as clearly as possible, the rationale for the study, the methods used and conclusions drawn. Sufficient detail should be presented so that a knowledgeable person could, in principle, replicate the study in order to check empirically on the outcome. However, it is not necessary that the report contains all individual results. An informed reader ought to be able to understand and develop a critique for the major details of the test, such as the underlying reasons for the study, the experimental design methods and execution, and the analyses and conclusions.

Special attention should be given to the following:

- a graphical presentation of the results;
- the specification and selection of subjects (see Note 1);
- the specification and selection of test material;
- general information about the system used to process the test material;
- details of the test configuration;
- the physical details of the listening environment and equipment, including the room dimensions and acoustic characteristics, the transducer types and placements, electrical equipment specification (see Note 2);
- the experimental design, training, instructions, experimental sequences, test procedures, data generation;
- the processing of data, including the details of descriptive and analytic inferential statistics;
- the detailed basis of all the conclusions that are drawn.

NOTE 1 – There is evidence that variations in the skill level of listening panels can influence the results of listening assessments. To facilitate further study of this factor experimenters are requested to report as much of the characteristics of their listening panels as possible. Relevant factors might include the age and gender composition of the panel.

NOTE 2 – Because there is some evidence that listening conditions, for example loudspeaker versus headphone reproduction, may influence the results of subjective assessments, experimenters are requested to explicitly report the listening conditions, and the type of reproduction equipment used in the experiments. If a combined statistical analysis of different transducer types is intended, it has to be checked whether such a combination of the results is possible (for example using ANOVA).

10.3 Presentation of the results

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

The results must be given together with the following information:

– description of the test materials;

– number of assessors;

– the overall mean score for all test items used in the experiment;

– only the mean scores and 95% confidence interval after post-screening of the observers, i.e. after eliminating those results according to the procedure given in Section 4.1.2 (post-screening).

Additionally, the results could also be presented in appropriate forms like histograms, medians or other.

10.4 Absolute grades

A presentation of the absolute mean grades, for the systems under test, the hidden reference, and anchor gives a good overview of the result. One should however keep in mind that this does not provide any information of the detailed statistical analysis. Consequently the observations are not independent and statistical analysis of these absolute grades will not lead to meaningful information and should not be done.

10.5 Significance level and confidence interval

The test report should provide the reader with information about the inherently statistical nature of all subjective data. Significance levels should be stated, as well as other details about statistical methods and outcomes, which will facilitate the understanding by the reader. Such details might include confidence intervals or error bars in graphs.

There is of course no “correct” significance level. However, the value 0.05 is traditionally chosen. It is, in principle, possible to use either a one-tailed or a two-tailed test depending on the hypothesis being tested.

References

EBU [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Draft EBU Recommendation, B/AIM 022 (Rev.8)/BMC 607rev, January.

EBU [2000b] EBU Report on the subjective listening tests of some commercial internet audio codecs. Document BPN 029, June.

SOULODRE, G. A. and LAVOIE, M. C. [September 1999] Subjective evaluation of large and small impairments in audio codecs, AES 17th International Conference, Florence, pp. 329-336.

Appendix 1
to Annex 1

Instructions to be given to subjects

The following is an example of the type of instructions that should be given to or read to the subjects in order to instruct them on how to perform the test.

1 Familiarization or training phase

The first step in the listening tests is to become familiar with the testing process. This phase is called a training phase and it precedes the formal evaluation phase.

The purpose of the training phase is to allow you, as an evaluator, to achieve two objectives as follows:

– Part A: to become familiar with all the sound excerpts under test and their quality level ranges; and

– Part B: to learn how to use the test equipment and the grading scale.

In Part A of the training phase you will be able to listen to all sound excerpts that have been selected for the tests in order to illustrate the whole range of possible qualities. The sound items, which you will listen to, will be more or less critical depending on the bit rate and other "conditions" used. Figure 3 shows the user interface. You may click on different buttons to listen to different sound excerpts including the reference excerpts. In this way you can learn to appreciate a range of different levels of quality for different programme items. The excerpts are grouped on the basis of common conditions. Three such groups are identified in this case. Each group includes 4 processed signals.

In Part B of the training phase you will learn to use the available playback and scoring equipment that will be used to evaluate the quality of the sound excerpts.

During the training phase you should be able to learn how you, as an individual, interpret the audible impairments in terms of the grading scale. You should not discuss your personal interpretation of the scale with the other subjects at any time during the training phase. However you are encouraged to explain artefacts to other subjects.

No grades given during the training phase will be taken into account in the true tests.

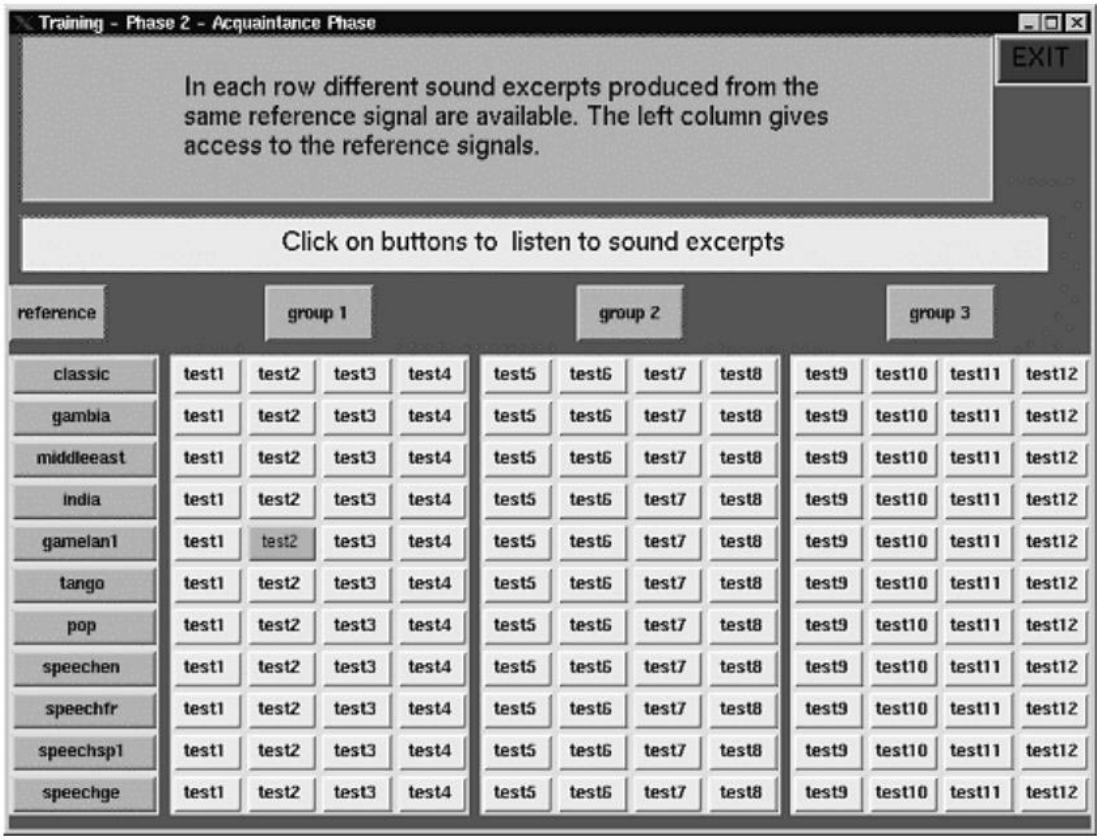
2 Blind grading phase

The purpose of the blind grading phase is to invite you to assign your grades using the quality scale. Your grades should reflect your subjective judgement of the quality level for each of the sound excerpts presented to you. Each trial will contain 11 signals to be graded. Each of the items is approximately 10 to 20 s long. You should listen to the reference and all the test conditions by clicking on the respective buttons. You may listen to the signals in any order, any number of times.

Use the slider for each signal to indicate your opinion of its quality. When you are satisfied with your grading of all signals you should click on the "register scores" button at the bottom of the screen.

FIGURE 3

Picture showing an example of a user interface for Part A of the training phase



1534-03

You will use the quality scale as given in Fig. 1 when assigning your grades.

The grading scale is continuous from "excellent" to "bad". A grade of 0 corresponds to the bottom of the "bad" category, while a grade of 100 corresponds to the top of the "excellent" category.

In evaluating the sound excerpts, please note that you should not necessarily give a grade in the "bad" category to the sound excerpt with the lowest quality in the test. However one or more excerpts must be given a grade of 100 because the unprocessed reference signal is included as one of the excerpts to be graded.

FIGURE 4

An example of the user interface used in the blind grading phase

1534-04

Appendix 2
to Annex 1

Guidance notes on user interface design

The following suggestions are made for those who might be considering:

a) producing systems for performing subjective tests according to the MUSHRA method,

b) performing such tests.

These suggestions are intended to increase the reliability of the results of tests and to facilitate the analysis of any irregularities that might be found during the processing of test scores.

The design of the user interface should be such that the chance of a subject assigning a score which does not accord their true intent is minimised. To this end, steps should be taken to ensure that it is clear from the user interface to which of the processed versions of a test item the subject is listening at a given time. This can be aided by careful choice of colours and brightness of on-screen indicators (clickable buttons, for example) to avoid potential difficulties should a subject not be sensitive to some colours.

It should also be ensured that the subject is only able to adjust the score assigned to the item currently being listened to. It has been observed that some subjects listen to two processed versions of an item, in succession, in order to assign a score to the first, not the last, that they hear. In this circumstance, it is possible that a mistake might be made (especially when a large number of on-screen controls are presented) and the score might be assigned to a signal other than the intended one. To try to reduce this possibility, it is suggested that the only control that is enabled at any one time is the one related to the signal currently being heard. Controls to assign scores to other signals, not currently being heard, should be disabled.