BBC

R&D White Paper

WHP 038

September 2002

The MUSHRA audio subjective test method

A.J. Mason

Research & Development
BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 038

The MUSHRA audio subjective test method

Andrew Mason

Abstract
The development of the audio subjective test method now known as MUSHRA took place in the EBU groups B/CASE and B/AIM leading eventually to an ITU-R Recommendation. The name is derived from the way in which the signals to be evaluated are presented and the way in which hidden references and anchors are included.

The design philosophy was to provide a less sensitive, but still reliable (repeatable, reproducible), test method that was less costly to implement than BS.1116. The way in which these intentions were realised is described.

This document was originally published at the EBU Specialised Meeting “Audio/Video Coding Technologies”, Geneva, 5th-6th September 2002.

White Papers are distributed freely on request.

Authorisation of the Chief Scientist is required for publication.

BBC Research & Development
White Paper WHP 038

The MUSHRA subjective test method

Andrew Mason

Introduction

Despite significant advances in objective test methods over recent years, the only way to find out the subjective quality of an audio signal is to ask people for their opinion. Quality of audio systems has, in general, increased as technical developments have been made. A subjective test method, BS.1116, was developed for assessing the subjective quality of high quality signals. It has been found to be very valuable.

The advent of the Internet has provoked a renewed interest amongst broadcasters in audio signals of lower quality, the restriction being imposed by the limited bandwidth of typical domestic connections. BS.1116, although well tried and tested, is not ideal for measuring the quality of this type of system. A need was seen for a new test method, one suited to "intermediate audio quality". An EBU project group was set up to develop such a method. Its first meeting was here on the 29th and 30th of August 1996. It passed the results of its work to the ITU and as a result we now have ITU-R Recommendation BS.1534.

This paper describes the main features of the test method and some of the reasoning behind the decisions made in its development.

Multiple Stimulus

The "MUS" part of MUSHRA stands for "multiple stimulus". This means that the subject is presented with all different processed versions of a test item at the same time. Not literally at the same time, but they are all available. This allows the subject to listen to one version of the item and then quickly switch to another version. In this way it is very easy to come to a decision about the relative quality of the different versions.

The original, unprocessed, version of the test item is also available to the subject. It is identified as the reference version. This guarantees that the listener knows how the item really should sound.

Hidden Reference and Anchors

The "HRA" part of MUSHRA stands for "hidden reference and anchors". Some of the multiple stimuli are pre-defined.

The original, unprocessed, version of each test item must appear as one of the versions to be graded by the subject. This is the hidden reference. It should always score quite highly - a low score would suggest that the listener is "unreliable".

Two other versions of each item have to be included. One of these is a low-pass filtered version with a bandwidth of 3.5kHz, the other is a low-pass filtered version with a bandwidth of 7kHz. A third, 10kHz low-pass filtered version is optional.

The purpose of the hidden reference and anchors is to ensure that the full range of the grading scale is used whatever the range of quality of the systems under test. Without the anchors it would be possible for a low quality system to get a higher grade than it should if the other systems are also low or lower quality. It works the other way too: a system could get a lower grade than it should because it is being tested with higher quality systems. The hidden reference ensures that the top of the scale is used, the 3.5kHz signal ensures that the lower end of the scale is used. The 7kHz anchor falls somewhere in the middle.

Although low-pass filtered versions might not exhibit impairments that are similar to the systems under test it was felt that specifying anchors such as “MPEG-1 Layer II at 128 kbit/s” would prove unreliable, firstly because there is too much flexibility in implementation, and, secondly, because it might be difficult to reproduce that process in a few year’s time. A low-pass filter is relatively easy to specify and implement unambiguously.

The grading scale

The subject is required to assign grades indicating his or her opinion of the quality of all the systems under test, the hidden reference, and the hidden anchors. The scale used is a numerical scale with descriptive terms associated with intervals on the scale. The scale runs from 100 to 0. The range from 100 to 80 is described as “excellent”, from 80 to 60 as “good”, from 60 to 40 as “fair”, from 40 to 20 as “poor”, and from 20 to 0 as “bad”.

The fact that there are only five descriptive terms does not mean that there is a restriction on the numerical values that a subject can assign to a version of an item. In this sense, the scale is continuous from 100 to 0 - any (integer) value can be used.

The test items

For reproducibility it would be a good idea to specify the audio programme material that should be used in all subjective tests. This idea has been considered by a number of different organisations over the years. The EBU produced the “SQAM” disc, which contains many different test items for use in subjective tests, and it has proved to be very popular. However, the EBU SQAM disc material was selected at a time when a lot of technical development work was being done on analogue-to-digital and digital-to-analogue converters. The kind of test material selected for revealing non-linearities in those devices is not always best suited to revealing quantisation distortion caused by bit rate reduction systems. This problem means that test material should always depend on the systems under evaluation. Every time that new systems are being tested, new test material should be sought.

Sometimes is it possible to create “pathological” test signals for particular systems. Whilst this kind of signal might be of academic interest, purely artificial signals are not recommended. Items should be representative of normal programme material. An artificial signal might be useful in suggesting an item - for example, a swept sine-wave might be catastrophic for some systems and this suggests that a US police siren could be a good test item.

A small panel of expert listeners, the selection panel, is normally given the task of selecting the set of items for testing. The usual method is to process many items with the systems under test, and then select, possibly in several stages, the set that is most revealing of artefacts in all systems. A range of programme types should be included and the set should not be unfairly biased in favour of, or against, particular systems.

Acoustic requirements

Although the audio quality that is being evaluated is “intermediate” the requirements relating to the environment used for the tests are high quality. The acoustic properties of the listening room are as stringent as they are in BS.1116. The reason for this is that it is the only way to ensure that the results are reproducible. If less stringent requirements were allowed then the results obtained from a noisier environment would not match those from a quiet environment. Different characteristics of inferior loudspeakers or headphones would also lead to results being difficult to reproduce.

The subjects

Expert listeners are preferred for subjective tests. The fact that the audio quality is not very high does not mean that expert listeners are not able to provide more consistent results more quickly. Inexpert listeners can become expert listeners simply by experience. Experience with, for example. MPEG audio coding, shows that listeners become attuned to artefacts that, once noticed, cannot be ignored again. Using expert listeners for subjective tests should mean that the tests produce the same results that inexpert listeners would eventually.

Of course, some listeners might be more expert than others. Post-screening of subjects is recommended for removing those results that show signs of sleep or of misunderstanding of instructions.

Running the tests

Instructions

The instructions given to the subject might be quite simple, but can significantly affect the way a subject performs the test. Example instructions are included in the recommendation. In the case where tests are being performed at several different sites it is more difficult to ensure that all subjects are instructed in the same way, especially if the instructions have to be translated.

There is still ambiguity in the descriptive terms used for quality. For example, “fair” can mean “good” in some circumstances, but a suitable replacement term was not easy to find.

Recording the sessions

Although not in the MUSHRA method itself, a useful addition is the recording of test sessions as they are conducted. Assuming that a PC is used to control the sessions it is relatively easy to make an audio and video recording that can solve problems found in the test results. A PAL version of the on-screen display, made by a VGA to PAL converter card, and a parallel feed of the audio signals can conveniently be recorded on a VHS recorder. A 4 hour tape, in long-play mode can record a whole day’s sessions for analysis in the case of suspected equipment or listener failure.

Analysing the results

A statistical analysis method is included in the MUSHRA method. Simply put, this is to calculate the mean grades and 95% confidence intervals. When presented with the results in this way, we can say, “The probability that the true value of a result lies within this confidence interval is 0.95.” To rely only on the mean value of grades is inappropriate and is only usually done because discarding all the other information produces a more favourable result.

The report

The MUSHRA method also specifies what should be included in the test report. Ideally this should be enough to enable another competent person to reproduce the tests. A good test report will convince a reader that the results are reliable, a poor one will leave too many questions unanswered.

Conclusions

The MUSHRA test method is the result of 6 years work by many people. It was devised to meet a need for a subjective test method that was appropriate to intermediate audio quality systems. It aims to result in tests that are repeatable and reproducible. The fundamental characteristics are:

1. multiple stimuli being offered to the subject;
2. one of the stimuli is a known reference;
3. one of the stimuli is a hidden reference;
4. other stimuli must include hidden anchors with clearly defined parameters.

The test method has been used by several different laboratories, notably for tests of internet audio codecs by the EBU project group B/AIM and is proving to be very effective.