THE QUEEN'S AWARDS
FOR ENTERPRISE
2001

BBC

R&D White Paper

WHP 118

September 2005

Cascaded audio coding

D.Marston *and* A.J. Mason

Research & Development
BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 118

Cascaded Audio Coding

David Marston, Andrew Mason

Abstract

Broadcasters have experienced significant problems with cascaded audio coding in the broadcast chain following the introduction of digital transmission. It has been found that cascading different codecs can result in an overall degradation in sound that many listeners find objectionable. A comprehensive investigation of this problem has been conducted by members of the EBU project group B/AIM.

This paper describes typical cascades of codecs found in radio broadcast chains, and aims to identify the most critical combinations. The intent is to guide broadcasters in deciding which codec combinations should be avoided to maximise sound quality.

The process initially involved deciding which were the most commonly used codec combinations used in digital radio. The next stage was to use objective assessment software, which gave an initial guide to the expected quality scores for each cascade. Finally, subjective tests involving trained listeners were performed to ensure more accurate and reliable results.

The resulting quality results were then analysed and conclusions drawn up to which cascades are best avoided.

This document was originally published in the Proceedings of the International Broadcasting Convention, September 2005.

White Papers are distributed freely on request.

Authorisation of the Chief Scientist is required for publication.

CASCADED AUDIO CODING

David Marston, Andrew Mason

BBC R&D, UK

ABSTRACT

Broadcasters have experienced significant problems with cascaded audio coding in the broadcast chain following the introduction of digital transmission. It has been found that cascading different codecs can result in an overall degradation in sound that many listeners find objectionable. A comprehensive investigation of this problem has been conducted by members of the EBU project group B/AIM.

This paper describes typical cascades of codecs found in radio broadcast chains, and aims to identify the most critical combinations. The intent is to guide broadcasters in deciding which codec combinations should be avoided to maximise sound quality.

The process initially involved deciding which were the most commonly used codec combinations used in digital radio. The next stage was to use objective assessment software, which gave an initial guide to the expected quality scores for each cascade. Finally, subjective tests involving trained listeners were performed to ensure more accurate and reliable results.

The resulting quality results were then analysed and conclusions drawn up to which cascades are best avoided.

INTRODUCTION

The production and broadcast of audio is a technically complex operation. The audio signal will typically pass through several distinct processes including recording, sending to the studio, post-production, and so on. Increasingly, people have been turning to bit-rate reduction to reduce the cost, or to increase the speed, of these processes. In isolation, the impact on audio quality of a single application of bit-rate reduction can appear negligible. However, in reality is that the effects of bit-rate reduction on all broadcast chain stages is far from negligible.

If each process removes all redundant audio information, or uses the signal to mask the noise being introduced, then the next process might have nothing left to remove, or will see previously introduced noise as signal to be used to mask more noise.

Whilst this is recognised, to a greater or lesser extent, there has still been little research into quantifying the effect of combining the huge numbers of bit-rate reduction codecs now available. It is more than 10 years since the ITU performed its extensive cascaded coding tests. Since then the number of commercially available codecs has increased enormously. One of the concerns raised by members of the EBU was that some of these codecs might interact badly with each other, producing much poorer sound quality than would have been expected. To fill this gap in our knowledge, the EBU embarked on an extensive test programme.

A typical broadcast chain was proposed containing 5 stages where bit-rate reduction could be used. The stages were called **acquisition**, **contribution**, **studio**, **distribution** and **emission**.

The intention was to process audio through all possible combinations of the 5 stages, and measure the subjective quality of the output. Subsequent analysis would be performed, comparing the subjective quality of the combinations. This would identify combinations whose performance was significantly worse, or better, that would be predicted from the performance of the codecs within them, in other combinations.

This report describes the process of identifying the chains of codecs to be tested and the process by which their subjective quality was measured. It then shows the results and the analysis done on them to find anomalous codec behaviour, if any.

A comparison is also made between objective and subjective quality measurements made during the tests.

INITIAL CASCADE SELECTION PROCESS

At each of the five stages there are several different audio codecs that can typically be used. Between 6 and 13 different codecs were originally identified as feasible at each stage, resulting in over 50,000 different cascade combinations. Therefore some refinement in codec selection was required to bring this number down to something more tractable for testing purposes.

The eventual selection of codecs for each stage is shown in Table 1.

Label	Codec	Bit-rate/kbps	Stereo mode	Position
O	PCM Linear	1536	Stereo	Acquisition
B	MPEG 1 Layer III	128	Joint stereo	
C	MiniDisc ATRAC	384	Stereo	
W	Windows Media 9	128	Stereo	
E	MPEG 1 Layer II	256	Stereo	Contribution
M	MPEG 1 Layer III	128	Joint stereo	
P	ADPCM	256	Stereo	
S	AAC	128	Stereo	
D	MPEG 1 Layer II	384	Stereo	Studio
E	MPEG 1 Layer II	256	Stereo	Distribution
F	MPEG 1 Layer II	256-12	Stereo	Emission
H	MPEG 1 Layer II	192-12	Joint stereo	
J	MPEG 1 Layer II	128-12	Joint stereo	

Table 1: Selected codecs

The emission codecs all had 12kbps removed from their normal bit-rate to simulate a typical DAB codec where some of the bits are reserved for data (FPAD and XPAD).

Only one codec option was chosen for the studio and distribution stages. These codec combinations give 48 possible configurations; however with the Windows Media acquisition codec, it was only likely to use the Layer II codec as a contribution codec. This brought the number of combinations down to a manageable 39.

SELECTION OF CASCADES FOR SUBJECTIVE TESTS

Subjective tests are used to measure, definitively, the human opinion of audio quality. Unfortunately, they are enormously labour-intensive. To make possible any prospect of subjectively testing cascaded codec quality, the number of cascades to be tested had to be reduced. Typically in a subjective test, the number of different stimuli presented to the listeners would be about 10. To use many more results in listener fatigue, and a difficulty in finding listeners!

It was decided to perform a screening process on the cascades using an objective quality assessment method. The chosen method was PEAQ [3]. Cascades with objective quality close to transparent, introducing imperceptible distortion, would not be subjectively tested - it would be assumed that they were adequate for broadcast use, even taking into account possible measurement errors.

Nine items of audio material were chosen for the PEAQ tests and subsequently for the subjective tests. They were selected in order to represent a good cross-section of types of broadcast material (speech, solo instruments, orchestral music) and to be items that would show differences between the codec chains. The list includes some old favourites, and is shown in Table 2.

Name	Description	Origin
accordion	Solo accordion music.	Swedish Radio
castanets	Castanets.	EBU SQAM CD
classic	Brass band music.	IRT
dialog	German male and female conversation.	T-Systems
harpsichord	Harpsichord playing an arpeggio.	EBU SQAM CD
orchestra	Classical music.	IRT
rea	Chris Rea.	Commercial CD
vega	Suzanne Vega, "Tom's Diner" a cappella.	Commercial CD
hockey	Commentary from ice hockey arena with crowd noise.	IRT

Table 2: Test items

Coding through the 39 cascades as chosen in the first stage, was performed by IRT and Radio France. The PEAQ objective quality assessment was performed by Radio France and the BBC. The results were in good agreement, although some technical difficulties were encountered.

Asynchronous operation of several of the hardware codecs gave varying time offsets between reference and coded items. This had to be corrected by precise sample rate conversion.

The results of the PEAQ measurements were a set of "objective diff grades" - ODGs. These are according to the well-known ITU 5-point impairment scale. An ODG of 0 means that there was no perceptible impairment with respect to the reference. An ODG of -4 means that there was a very annoying difference.

It should be pointed out that PEAQ was designed originally to perform measurements according to the ITU-R BS.1116 [1] scale. Early implementations were found to be unreliable when presented with large impairments. It was hoped that the poorer quality cascades could be useful in checking the performance of PEAQ at this lower end in quality.

As a result of the objective measurements the three cascades with ODGs between 0 and -1 were eliminated from subjective testing. This left 36 cascades.

SUBJECTIVE TESTING

There are two subjective methods that are commonly regarded as standard: MUSHRA[4] and ITU-R BS.1116. MUSHRA gives the listeners the option of selecting an absolute quality score from several coded versions of the original audio. ITU-R BS.1116 compares the original with a coded version, and an impairment score is given. MUSHRA is better suited to poorer audio quality, which is what was expected from these tests, so was chosen. It also has the advantage of being faster to perform as several stimuli are presented at once.

MUSHRA

In the MUSHRA test the listener is presented with several audio stimuli. The first is the reference, which is the original uncoded audio. The remainder are the test stimuli to which the listener must give a score between 0 and 100 depending upon their opinion of the quality. The scale is given this range of quality categories: “excellent” (100-80), “good” (80-60), “fair” (60-40), “poor” (40-20) and “bad” (20-0).

Amongst the test stimuli there are 3 anchors which must appear: a hidden reference (i.e. the identical audio to the reference), a 3.5kHz low pass filtered version of the reference and either a 7kHz or 10kHz low pass filtered version. In these tests the 10kHz anchor was chosen. The listener must attempt to identify the hidden reference and score it as 100. The test stimuli must be randomised in order, so the listener has no clues to their identity.

Each test item (clarinet, harpsichord, etc.) is presented separately, so the listener must complete each test item before going onto the next one. In these tests there were 9 test items, so it was recommended the listeners took some breaks in between some of them to reduce fatigue.

Software

The MUSHRA listening test software used for these tests was developed by Fraunhofer, and it provided the user with a series of sliders for scoring, and buttons for playing the various test stimuli.

Listening Test Environment

Listening was carried out using headphones for both convenience and to give more consistent conditions than loudspeakers in listening rooms. The tests were carried out with Stax electrostatic headphones in a quiet room, where equipment noise was not audible.

Allocation of Cascades to Sites

The objective test results were used as a guide to the allocation of the cascades to each of the 5 sites. Each site would get an even spread of cascades, so every fifth cascade in the ordered (according to the ODGs) list would be given to each site. To test consistency between sites, three cascades (a high, medium and low scoring cascade) were given to all five sites. This would allow a comparison of the sites' overall performance, and help assess the results.

Filtering of Subjective Results

Listening tests rely on reliable and well trained listeners to give accurate results. It is not always possible to get experienced expert listeners to volunteer, so it was important to give all the listeners a good training session before the tests. They were exposed to all the test items used, and a selection of the coded versions.

To aim for a minimum of 15 valid listeners per site, more than this number was used to allow for poorly performing listeners. To decide which listeners achieved an acceptable level of performance, the ability to identify the difference between the hidden reference and the 10kHz low-pass filtered anchor was used. Any listeners who could not tell the difference between these two stimuli on most of the test items were rejected. Most of the listeners who fell into this category were over the age of 50, an age when high frequency sensitivity begins to reduce.

ANALYSIS OF RESULTS

The subjective scores for each of the 36 cascades that were tested are shown in Figure 1. This plot shows the average score for each cascade over all test items, with 95% confidence intervals shown. The first three cascades are the cascades common to all five test sites. The remaining cascades are ordered by their objective test scores.

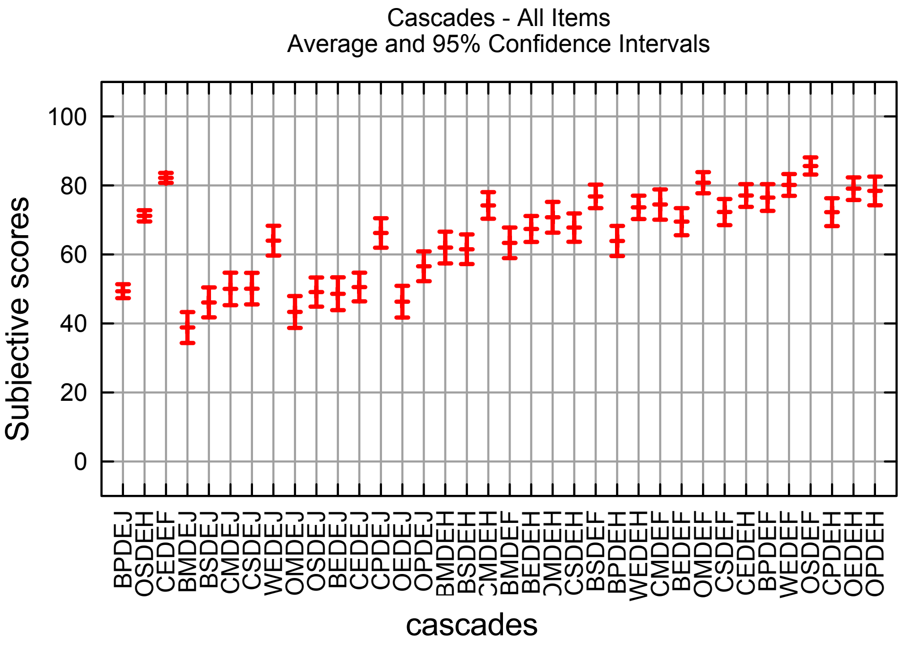

Figure 1: Subjective scores of the 36 cascades.

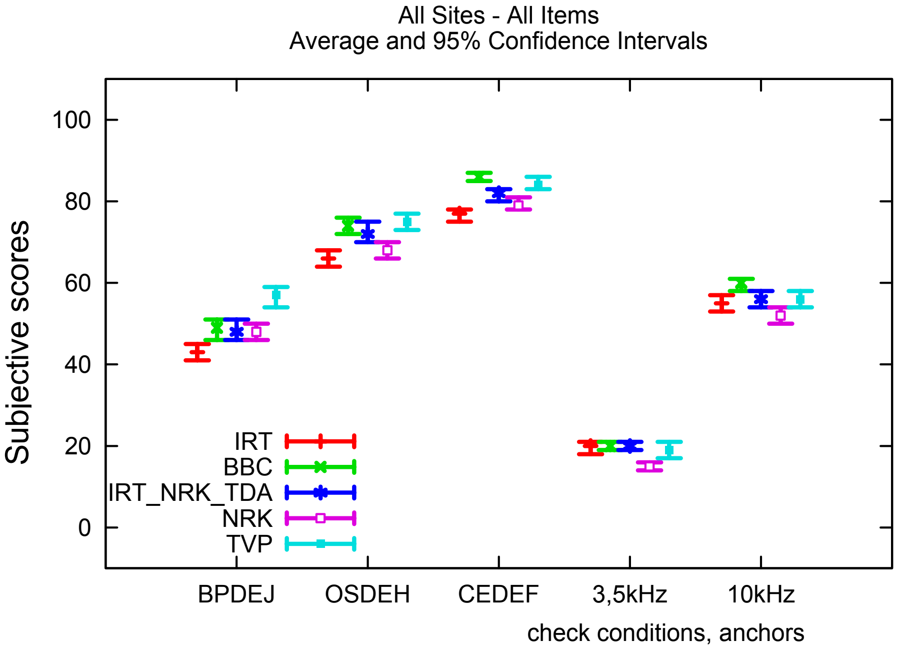

Figure 2: Cascades common across sites plus the anchors

Site Dependencies

Figure 2 shows how each of the five sites performed with the three common cascades

(BPDEJ, OSDEH, CEDEF), and the two anchors. It can be seen that generally IRT scored the lowest, while TVP scored the highest. These differences are significant as the confidence intervals are not overlapping. This pattern is reflected in figure 1 where every 5th point follows the same shape.

PREDICTION ANALYSIS

It is not immediately obvious from the set of MUSHRA scores whether some cascades are better or worse than one would expect or not. An analytical method of making a prediction of a chain's performance, based on data describing the constituent codecs' performance in other combinations is required.

The method used has two stages. In the first stage, a MUSHRA score is assumed to be the sum of a contribution from each codec in the chain. On that assumption is based a calculation of an "impairment coefficient" for each codec in the tests.

Mathematically this can be expressed as $A*C = M$ where C is a column vector of all codec impairment coefficients, M is column vector of the MUSHRA grades for all chains, and A is a rectangular matrix of 1s and 0s, each row corresponding to one chain, containing a 1 at 5 locations to pick out the codec coefficients making up the chain in C. Implicit in this model is also the assumption that the order in which the codecs are applied makes no difference.

The aim then is to find the minimum norm least square solution for C. The mathematical software tool called "scilab" [2] was used to perform the calculation. To calculate the values of C it was simple necessary to enter the data for A and M and set $C = lsq(A, M)$.

The second stage is to calculate the predictions. This is simply $A*C$. The prediction errors are then $A*C - M$.

A problem was apparent in the ranking of codecs by their calculated coefficients. The allocation of chains to sites, combined with the site dependency of the scoring meant that one codec appeared to be better then the original. It was therefore difficult to rely on the predictions from this.

The same predictive process tried on the objective test results showed the ranking that one would expect, and did not show any significant anomalies in the performance of the chains tested.

MUSHRA VERSUS PEAQ

The test procedure described above included objective and subjective tests on the same codec chains. One aim of this was to allow for verification of the objective test method, particularly at the lower quality range. Before making any comparison, it is important to note that the objective quality measurement produced scores according to the ITU 5-point impairment scale, whereas the MUSHRA method used for the subjective tests produces results on an absolute quality scale. Let's look at the results and then come back to this point.

Figure 3 shows a scatter plot of the subjective quality score and objective quality score for each of the 36 chains. A general trend is quite apparent as indicated by the straight line. The site dependencies account for much of the scatter.

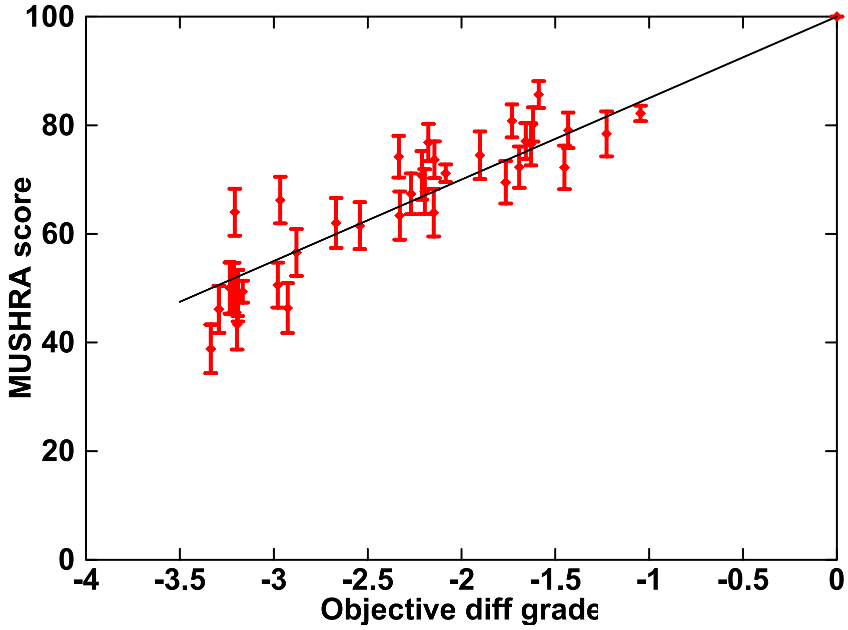

Figure 3: PEAQ versus subjective scores for all cascades

However, if the question is asked, "Is there a mapping from MUSRHA scores to PEAQ scores", it must be answered with care. Inspection of the graphs suggests that a MUSHRA score of around 50 (a quality of “fair”) corresponds to a PEAQ score of corresponds to something between "Annoying" and "Very annoying". Clearly, from the normal English usage of the words, this is not appropriate. This raises the question as to whether the inconsistency is due to deficiencies in the test methods or not.

The test method specified in ITU-R BS.1116 is described intended for measuring small impairments. It is designed to be sensitive, and listeners would be asked to assess the differences between the reference and the processed signal. ITU-R BS.1387 (PEAQ) was designed to be able to be used in its place. The MUSHRA test method is designed to cover a wide range of audio qualities, hence the presence of anchors (the bandwidth-limited signals) at the various qualities specified. It also asks listeners to grade the processed signals in an absolute way, not according to the differences - the reference is there simply so that listeners know what the signal should really sound like.

Bearing this in mind, it is not therefore sensible to expect a usable mapping between MUSHRA scores and PEAQ ODGs. That there appears to be some correlation between the two is nice, but it would be a mistake to assign great importance to it.

To verify the performance of PEAQ at lower qualities would have required the use of ITU-R BS.1116 for the subjective tests. This was ruled out for the reasons given earlier.

CONCLUSIONS

An extensive, thorough, and time-consuming investigation has been conducted by members of the EBU B/AIM project group into cascaded audio coding. A model of a broadcast chain consisting of 5 cascaded codecs was assumed. From the thousands of possible combinations of codecs, a subset of the more likely ones was tested for audio performance using objective and subjective methods.

Objective testing using PEAQ was successfully employed to reduce the number of combinations to be subjectively tested. The subjective testing was performed using the

MUSHRA test method, with the subset of codec combinations being divided amongst a small number of test laboratories. Some codecs were tested by all sites for comparison purposes.

The results clearly show that the cumulative effect of cascaded audio coding can be highly detrimental to audio quality, even when each stage in the chain makes only a small reduction in quality.

The comparison of objective and subjective results showed a good correlation between scores. Caution should be exercised here because the scales and descriptive terms associated with the two test methods used are quite different.

The objective and subjective test results were both analysed to try to identify codec performance that was significantly better or significantly worse than expected. It was found that none of the combinations showed any unusual behaviour. This should simplify the selection process for users of low bit rate coding - it implies that choosing the best codecs will give the best results.

Refinements to the MUSHRA method

The current MUSHRA method states that two anchors of low-pass filtered audio are required. It was found that this type of audio sounds very different from low bit-rate or cascade coded audio with a large bandwidth, thus it becomes difficult to make comparisons. A possible solution would be to use some sort of sub-band coded anchor, possibly based around a modified MPEG coder, where the artefacts are of a similar nature to the tested codecs. During the original design of the test method, the idea of a specified anchor codec for MUSHRA had been suggested, but the difficultly and cost of maintaining it was deemed to be prohibitive.

The differences between the sites posed problems in the analysis, the reasons not fully understood. However it was considered that younger listeners tended to be more 'generous' with their scoring. Therefore trying to ensure a reasonable age spread of listeners should even out scoring; bearing in mind that older listeners often struggle to hear high frequencies.

Interpretation of the meanings of 'excellent' down to 'bad' may differ between individuals and also languages and cultures. Some method of unifying or clarifying these definitions may have to be made.

ACKNOWLEGMENTS

Thanks are due to the members of the EBU B/AIM group for their essential contribution to this work.

REFERENCES

[1] International Telecommunications Union Recommendation ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", ITU

[2] "Scilab - A Free Scientific Software Package", Institut National de Recherche en Informatique et en Automatique, http://scilabsoft.inria.fr/

[3] International Telecommunications Union Recommendation ITU-R BS.1387-1, "Method for objective measurements of perceived audio quality", ITU

[4] Method for the subjective assessment of intermediate quality level of coding systems", U-R Recommendation BS.1534 (June 2001)