 ISO/IEC JTC 1/SC 29/WG 6 N341

**ISO/IEC JTC 1/SC 29/WG 6**

**MPEG Audio Coding Convenorship: DE**

**Document type:** General

**Title**: Rules for core experiments for audio coding for machines

**Status:** Approved

**Date of document:** 2025-04-04

**Source:** ISO/IEC JTC 1/SC 29/WG 6

**Expected action:** None

**Action due date:** None

**No. of pages: 13** (including cover page)

**Email of Convenor:** Thomas.Sporer @ idmt.fraunhofer.de

**Committee URL:** https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg6

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG6, MPEG AUDIO CODING**

**ISO/IEC JTC1/SC29/WG6 N341**

**April 2025, Virtual**

|  |  |
| --- | --- |
| **Title** | **Rules for core experiments for audio coding for machines** |
| **Source** | **WG 6, MPEG Audio Coding** |
| **Status** | **Approved** |
| **Serial Number** | **25013** |

# Scope

This document describes the core experiment procedures for the MPEG ACoM work. The recommendations in this document are based on previous MPEG Audio CE process documents (N1375, N1504, N1748, N5722, N7140, N10664 and N13634).

# Foreword

There are two different type of CE experiments: improvement of the performance of the FoM as defined for CfP, and improvement concerning additional requirements.

* Compared to RMx improvement of figure of merit by more than 10% either considering the average compression rate or improvement of the worst-case performance by more than 10% without reducing the average performance.
* Improvement of additional requirements include smaller access unit, less computational complexity without reduction of coding efficiency. CE accepted if at least 10% improvement compared to RMx.

**Definitions**

The following definitions apply to this document:

* **Working Draft** (WD)refers to the textual description of the technology under development. Note that in the final phases of the standardization process the text is referred to as CD, DIS, FDIS and IS.
* **Reference Software** (RefSoft) refers to an implementation in C/C++ source code (defined below) of the technology specified in the text. This includes both the Reference Software Encoder (RefSoft Encoder) and the Reference Software Decoder/Renderer (RefSoft Decoder/Renderer).
* **Reference Model** (RM) refers to the technology under development, which is expressed in both the WD and the RefSoft. Any addition or change to the RM implies an addition or change to both the WD and the RefSoft.
* **C/C++ source code** refers to ISO/IEC 14882:2017 (informally known as C++17) source code with C++ compatible header files.

**Introduction**

Experiments shall not affect more modules in the RM than necessary. If there is doubt whether this is the case, WG 6 Audio may request a more detailed check as a means to understand the effect of smaller-scale changes.

A core experiment always starts at one MPEG meeting with the submission of a proposal and the naming of the parties doing independent coding and/or volunteering to do part of the evaluation. It can span the time between more than two MPEG meetings.

The MPEG process for developing of standards is documented in [1]. New MPEG delegates are urged to review this document.

# The Core Experiment process

To include a new tool or algorithm into the MPEG ACoM RM, the following three steps are necessary. Only after the completion of all three does a new tool becomes part of the RM. The three steps are:

1. Proposal

2. Check

3. Integration

### 1 Proposal

In the proposal stage, an internal verification and evaluation must be done by the proponent(s). It is assumed that the proponent already has code implementing the proposed new tool or changed tool.

A proposal must be submitted as an MPEG contribution document and must contain:

1. A one page “executive summary”
2. Information as to whether the new tool is planned to be an addition to the RM or if it is planned to replace a tool in the RM.
3. A technical description which is accurate enough that the main function can be duplicated by someone skilled in the art (this does not mean bit-identical replication of results).
4. The WD changes (bitstream syntax and sematics and a description of the decoding process) which are necessary to incorporate the proposed technology.
5. A complexity evaluation giving estimates of ROM, RAM and signal processing functions (e.g. weighted MOPS) necessary to implement the tool. See Annex B.
6. Evidence for the merit of the new tool (for example subjective test results), see Annex B and C. If CE is to add support for a new functionality, then proposal must reference the MPEG ACoM requirement that the new functionality satisfies, use cases that this might apply to, and provide evidence of why the proposed technology is the best solution.

A CE shall be conducted using the Reference Software Encoder (RefSoft Encoder) and the Reference Software Decoder/Renderer (RefSoft Decoder/Renderer).

Manual tuning (i.e. per-item tuning) of the encoder used in the process of preparation of test material is explicitly prohibited.

If there is a conflict between proposals in which the tool of one CE affects the tool of another CE (i.e. a first CE changes the RM environment in which the second CE must be evaluated), the CE whose proponent was first to present a complete proposal shall have precedence.

If there is a conflict between proposals in which both CEs target the same functionality (i.e. two versions of the same tool), proponents should attempt to forge a unified proposal. If this is not possible, a more rigorous evaluation must be done. In this case, objective parameters (delay, complexity) should be part of the evaluation.

### 2 Check

The Check phase ensures that the proposal is implemented, tested and reviewed by at least one company in addition to what was provided by the proponent in the Proposal phase. The Check phase information must be submitted as an MPEG contribution document.

Optionally, if there is unanimous agreement of WG 6 experts that the Proposal is of sufficient merit, then the verification of implementation and verification of performance can be skipped, and the Proposal can proceed directly to Integration.

**Verify Implementation**

This is a check on the software implementation of the proposal. At least one check by an independent company must be performed. (It is assumed that one check is provided by the proponent in the Proposal phase.) It is acceptable if this is using the same software implementation, however an independent implementation is preferred. At the Check phase, proponent core experiment encoder and decoder/renderer software (i.e. extensions to the RefSoft Encoder and Decoder/Renderer) can be made available as an executable or as object modules with a suitable interface to the RefSoft Encoder and Decoder/Renderer. However, the successful completion of the core experiment requires that source code be delivered and integrated into the RefSoft Encoder and Decoder/Renderer, as described in Section 3 of this document.

**Verify Performance**

Assessment of performance (e.g. subjective tests) should be done by at least one lab. (It is assumed that the proponent provided one assessment of performance in the Proposal phase.) See Annex B for details concerning conducting subjective tests and measuring complexity and Annex C for details concerning statistical calculations. The requirement for an additional subjective test is dropped if an independent commercial test lab conducted the Proposal phase evaluation.

The result of the Check phase is a report on the technique, its merits and associated test results. The result should take into account the improvement (if any), theoretical considerations and the complexity of the feature.

In the case that a subjective test is part of the Check phase, material used in the subjective test must be made available at a reasonable interval (to be determined at each MPEG meeting) prior to the MPEG meeting where the results are presented, such that other members of MPEG have the opportunity to do informal evaluation.

If the result of the Check phase is not consistent amongst the companies doing the review, a report should present the different results, and WG 6 Audio may request an additional check.

The acceptance of a core experiment is based on the guidelines described in Annex D.

### 3 Integration

After the success of a core experiment, encoder and decoder/renderer components for the new tool must be integrated into the WD and the RefSoft. Source code shall carry the Copyright Header as shown in WG6 N337.

After integration, it is expected that the RefSoft Decoder/Renderer shall be able to decode the CE bitstreams to produce the bit-exact waveforms used in the CE evaluation.

For the RefSoft Encoder, the proponent must provide

* The RefSoft Encoder must include source code that implements the described extraction algorithms. Source code can be in C, C++, Python or Matlab that is able to be integrated into the RefSoft Encoder to make a working encoding system. Note that this need not be the exact algorithm used in the proponent’s CE. *However, subsequent Core Experiment work shall use the proponent-submitted MPEG ACoM “encoder” algorithm and corresponding submitted source code as the “baseline” system in CE performance comparisons, against which CE technology is evaluated.*

Satisfactory completion of the Integration phase concludes the Core Experiment. This integration is typically concluded in the AHG period following the MPEG meeting at which the Check phase was successfully completed.

**References**

1. N6877, MPEG 101, issued at 71st MPEG meeting.
2. ITU-R Recommendation BS.1284-2, General methods for the subjective assessment of sound quality, 2019, Intern. Telecom Union, Geneva, Switzerland (specifically, A-B Comparison)
3. ITU-R Recommendation BS.2132, Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli Without a Given Reference, 2019, Intern. Telecom Union, Geneva, Switzerland.

# Annex A: Core Experiment Process Flow Chart

Diagram

Description automatically generated

**Annex B: Performance Measurement Procedures**

1. **Procedure for subjective quality tests**

* A minimum of 8 subjects is required at each test site.
* The test material that might be used in the core experiment process is identified in Annex A of WG6 N339.
* If tools evaluated with the CE process are applicable to a specific type of test material, the regular set of items can be restricted to a subset, or new material can be added. WG 6 Audio must agree this to on a per-CE basis.
* Subjects must participate in a training phase prior to the actual test.
* Results should be presented for each item (averaged over all listeners) and per systems (averaged over all listeners and all items). In addition, results may be presented for pooling over all test sites and all operating points. In any case, subjective test results from all test sites should be considered.
* The CE technology may be assessed using differential scores and a single-sided t-test (see Annex C) if there is only a single CE configuration to be compared against a single Reference configuration. Single-sided test results should be for data pooled over all listeners and all test items.
* Subjective test methodologies can be
  + A-B Comparative (BS.1284-2) [2] (recommended when assessing a single CE proposal against the baseline system).
  + One of the methods in BS.2132 [3] (recommended for the assessment of newly proposed functionalities for which no baseline system performance is available as a benchmark, or for assessing absolute levels of quality, or for assessing the size of subjective differences as opposed to a mere ranking).
* If Comparative, A-B methodology, then
  + Score data should be analyzed and presented as improvement (or degradation) relative to the RM system. Better performance at a 95% level of significance is indicated by Confidence Interval (CI) not overlapping with 0 (i.e. the score for “no preference” or “same as”).
* If BS.2132 then
  + If there is a baseline system, score data should be analyzed and presented as both absolute scores and diff-scores relative to the baseline RM system (i.e. (RM+CE) – RM). Differences are to be computed on a per-listener-response basis. Better performance at a 95% level of significance is indicated by CI of CE proposal not overlapping with CI of baseline system. Additional statistical analysis procedures can be found in BS.2132.

1. **Procedure for complexity reduction or functionality addition with possible quality degradation**

* Use the same procedure for subjective quality assessment as in Section 1, above.
* Measure CE reduction in computational complexity using weighted MOPS (e.g. 1 instruction per MAC, 10 instructions per square root or log operation). Report:
  + Complexity of CE technology
  + Potential for parallization of computations
  + Total complexity of RM+CE
  + Ratio of (RM+CE)/RM complexity (i.e. percent change in decoder complexity when incorporating CE).
* Measure CE reduction in table size (i.e. ROM size) in Kbytes or Mbyes. Report:
  + Complexity of CE technology
  + Total complexity of RM+CE
  + Ratio of (RM+CE)/RM complexity (i.e. percent change in decoder complexity when incorporating CE).
* The proponent can measure functionality in any manner judged most appropriate.

1. **Procedure for enhanced functionality or increased compression producing bit-identical output**

* Report the functionality and the complexity as in Section 2, above.
* At least one additional independent checks on functionality or compression and identical reconstruction. (It is assumed that the proponent provided one such assessment in the Proposal phase.)

## Annex C: Statistical Calculations

## Confidence Interval on the Mean Score

A 95% confidence interval on the mean of a statistic is calculated as

,

where is the Student t statistic with single-sided tail probability ofand *n-1* degrees of freedom,

*μ* is the average

,

and *σ* is the standard deviation

,

where *N* is the sample size (e.g. number of listeners) and *xk* denotes the individual sample values (e.g. individual listener score).

Note that index *k* in xk enumerates the number of listeners and test items and is appropriate for estimates that are pooled across all listeners and all test items. If estimates per test item are desired, xk could be expressed as  where  is the subject (i.e. listener) index and  is the item index.

Note further that  converges to 1.96 as n goes to infinity, so that for large n, the t-distribution converges to the Normal distribution. It is within 2% of 1.96 for n > 60.

## Confidence Interval on the Mean of Diff-Scores

If it is desired to analyze the difference between scores on a per-listener-response basis (e.g. between a system under test and the reference system), then statistic is replaced by



where

 are the scores for the reference system

 are the scores for the system under test

A 95% confidence interval on the mean of the difference statistic is calculated as indicated above.

## Single-Sided Test

A single sided test on the difference between scores can be applied in order to determine whether the system under test is better than the reference system, in a statistical sense. This tests whether the assumption that there is no difference between the reference system and the system under test (the so called Null hypothesis) is rejected. A criterion is set such that there is an error of 5% (Type I error) that we falsely rejected this assumption. Given a 5% Type I error probability, the critical value is  times the standard deviation for the t-distribution, where converges to 1.64 for large n.

The testing procedure is as follows:

- calculate the critical value 

- test whether 

There are only two possible outcomes:

**Case 1**: . The Null Hypothesis is rejected and it is concluded that the system under test is better than the reference at the one-sided 95% level of significance

**Case 2**: . The Null Hypothesis is not rejected so that we cannot conclude that the system under test is better than the reference.

## Student t and Normal distributions

The statistical calculations above are given for both Student t and Normal distributions. Generally, [Student's *t* distribution](http://en.wikipedia.org/wiki/Student%27s_t-distribution) must be used to calculate confidence intervals when the number of observations is small (i.e. n < 60). This would be appropriate when computing per-item confidence intervals where the number of listeners is small (e.g. 8). In the case that the Normal distribution is used instead of Student's *t* distribution to calculate confidence intervals, this must be clearly stated when test results are reported.

## Annex D: Evaluation of Performance Information

As part of the Check phase approval process, WG 6 Audio will classify CEs into one of the following three categories and apply a separate process for each of these categories:

* **Sensible engineering**  
  Proposals in this category typically do not provide any gain in terms of quality, bitrate or complexity. The objective of such proposals is to resolve bugs and inconsistencies in text and reference SW or provide better alignment with other modules. Proposals in this category should be discussed amongst the experts in WG 6 Audio after which a decision based on technical arguments is made on whether to adopt, reject or possibly reconsider the proposal at a future MPEG meeting.
* **Performance gain in quality, bitrate or complexity**  
  Proposals in this category target an improvement in quality, reduction in bitrate or reduction in complexity. Figure 1 and the following discussion give guidance as to how to proceed with this type of proposal.

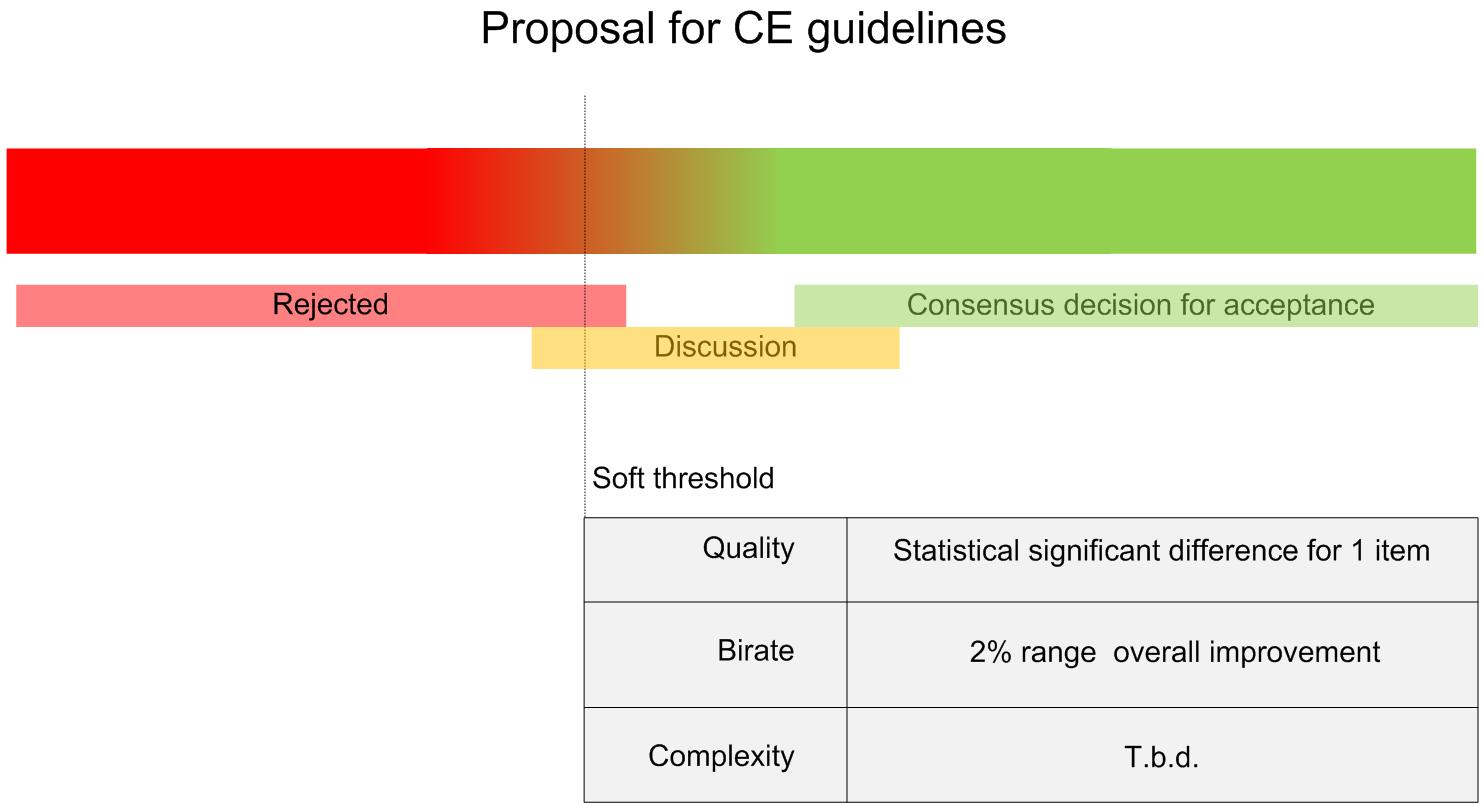


Figure 1 – Illustration of CE “Soft Threshold” concept.

As part of the CE requirements, each CE needs to provide clear information on its improvement in terms of subjective test results, overall bitrate reduction and/or overall complexity reduction. These data will rank the proposals with respect to the soft threshold illustrated in Figure 1. The soft threshold for quality, bitrate and complexity is given in the following table:

|  |  |
| --- | --- |
| **Parameter** | **Soft Threshold** |
| Quality | Statistically better for 1 item, none worse |
| Bitrate | 5% overall improvement |
| Complexity | 5% overall improvement |

A discussion of these thresholds follows:

* **Quality**  
  Statistically significantly better for at least 1 item and no item worse with respect to the current WD status. This may be determined using absolute subjective or differential subjective results using two-sided confidence intervals. “Better for at least 1 item” means that there is at least one item for which the confidence interval for the (RM+CE) mean score is higher than and does not overlap with the confidence interval for the corresponding RM mean score. “No item worse” means that there is no item for which the confidence interval for the (RM+CE) mean score is lower than and does not overlap with the confidence interval for the corresponding RM mean score.

The notion of Just Noticeable Difference (JND) for a given subjective test methodology and test scale should be discussed and taken into consideration. CE technology whose subjective improvement does not surpass the JND clearly does not provide a perceptible difference.

* **Bitrate**  
  Although proposals targeting only bitrate reduction typically accomplish this without any change in decoded output (at least when issues of bit budget and bit allocation are ignored), in the end the objective is that bitrate savings give rise to an increase in sound quality. With this in mind, it is observed that differences in bitrate below 5% are too low to result in a noticeable difference in a subjective test. It is also recognized that a noticeable gain in quality could of course become apparent by combining different proposals that do not show a difference in quality when tested individually. It is however more likely that a noticeable gain in subjective quality can only be achieved when combining proposals that show a gain in bit-rate in the order of 5%. Proposals with a gain well below 3% will in the end likely have no contribution to the quality improvement. For this reason, the soft threshold is set to 5% for proposals targeting bitrate reduction.
* **Complexity**  
  In some application scenarios, issues of complexity are of primary concern for the decoder, and typically much less of a concern for the encoder. Complexity enters the discussion in two ways. Either a proposal has as its main merit the reduction of complexity, or a proposal has another merit of e.g. quality improvements at a cost of a certain additional complexity. Either way, the complexity, be it computational complexity or RAM/ROM requirements, needs to be presented in relation to the entire complexity of the decoder system. Furthermore, it is imperative that the complexity measurements are done in a scientific and rigorous manner e.g. by counting WMOPS, KB of table size, or other parameters that the CE proponent feels are relevant for measuring complexity. With this in mind, measurements of decoder execution time on a specific platform cannot be considered to constitute sufficient evidence of any complexity increase or reduction. It is further noted that providing the complexity numbers in relation to functionally comparable technology in the marketplace would allow experts in the group to better judge the proposal. A soft threshold is set at 5% overall reduction in complexity.

For proposals with a merit below the soft threshold:

The larger the distance below the soft threshold, the less likely it should be to gain the support of audio experts, subject to consideration of all other evidence of merit.

For proposals with a merit that is above the soft threshold:

The larger the distance above the soft threshold, the more likely it should be to gain the support of audio experts, subject to consideration of all other evidence of merit.

* **Alternative operating point**

Proposals in this category target a significantly different operating point than the existing technology, for example supporting an operating parameter of lower bitrate, -computational complexity or -memory. Simultaneously, any negative impact on other parameters of the standard should be much less pronounced compared to the improvements in the target operating parameter. For example, for a low computational complexity proposal, any reduction of quality, increase in bitrate, increase in memory, or increase in implementation complexity should be within a reasonable range. This is subject to discussion among and consensus agreement of WG6 experts. When considering alternative operating points, it is desirable to provide just a few (perhaps just two) operating points, e.g., to provide “high quality” and “low complexity” operating points.

Furthermore, the new technology should be well integrated within the stage and with other stages. It should allow combination with the other stages in a meaningful way. If other stages have similar alternative operating points, the new technology should work well with the corresponding mode of the other stage. This integration shall be demonstrated by the proponent.

Finally, the proponent shall motivate why the degradation of some evaluation measures (i.e., operating parameters) is acceptable given the significant improvements of others.