

CODEC
Internet Draft
Intended status: Informational
Expires: December 2011

C. Hoene
Universitaet Tuebingen
June 3, 2011

Measuring the Quality of an Internet Interactive Audio Codec
draft-hoene-codec-quality-01.txt

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on June 3, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

The quality of a codec has to be measured by multiple parameters such as audio quality, speech quality, algorithmic efficiency, latency, coding rates and their respective tradeoffs. During standardization, codecs are tested and evaluated multiple times to ensure a high quality outcome.

As the upcoming Internet codec is likely to have unique features, there is a need to develop new quality testing procedures to measure these features. Thus, this draft reviews existing methods on how to measure a codec's qualities, proposes a couple of new methods, and gives suggestions which may be used for testing the Internet Interactive Audio Codec (IIAC).

This document is work in progress.

Conventions used in this document

In this document, equations are written in Latex syntax. An equation starts with a dollar sign and ends with a dollar sign. The text in between is an equation following the notation of Latex Version 2e. In the PDF version of this document, as a courtesy to its readers, all Latex equations are already rendered.

Table of Contents

Conventions used in this document	2
1. Introduction	4
2. Optimization Goal	6
3. Measuring Speech and Audio Quality	7
3.1. Formal Subjective Tests	7
3.1.1. ITU-R Recommendation BS.1116-1	7
3.1.2. ITU-R Recommendation BS.1534-1 (MUSHRA)	8
3.1.3. ITU-T Recommendation P.800	8
3.1.4. ITU-T Recommendation P.805	8
3.1.5. ITU-T Recommendation P.880	9
3.1.6. Formal Methods Used for Codec Testing at the ITU ...	9
3.2. Informal Subjective Tests	9
3.3. Interview and Survey Tests	9
3.4. Web-based Testing	10
3.5. Call Length and Conversational Quality	10
3.6. Field Studies	12
3.7. Objective Tests.....	13
3.7.1. ITU-R Recommendation BS.1387-1	14
3.7.2. ITU-T Recommendation P.862	14
3.7.3. ITU-T Draft P.OLQA	15

4. Measuring Complexity	15
4.1. ITU-T Approaches to Measuring Algorithmic Efficiency ...	15
4.2. Software Profiling	17
4.3. Cycle Accurate Simulation	18
4.4. Typical run time environments	19
5. Measuring Latency	19
5.1. ITU-T Recommendation G.114	20
5.2. Discussion	20
6. Measuring Bit and Frame Rates	21
7. Codec Testing Procedures Used by Other SDOs	22
7.1. ITU-T Recommendation P.830	22
7.2. Testing procedure for the ITU-T G.719	24
8. Transmission Channel	25
8.1. ITU-T G.1050: Network Model for Evaluating Multimedia Transmission Performance over IP (11/2007)	26
8.2. Draft G.1050 / TIA-921B	27
8.3. Delay and Throughput Distributions on the Global Internet	27
8.4. Transmission Variability on the Internet	30
8.5. The Effects of Transport Protocols	30
8.6. The Effect of Jitter Buffers and FEC	33
8.7. Discussion	33
9. Usage Scenarios	34
9.1. Point-to-point Calls (VoIP)	34
9.2. High Quality Interactive Audio Transmissions (AoIP)	35
9.3. High Quality Teleconferencing	35
9.4. Interconnecting to Legacy PSTN and VoIP (Convergence) ..	36
9.5. Music streaming.....	36
9.6. Ensemble Performances over a Network	36
9.7. Push-to-talk like Services (PTT)	37
9.8. Discussion	38
10. Recommendations for Testing the IIAC	38
10.1. During Codec Development	38
10.2. Characterization Phase	39
10.2.1. Methodology	39
10.2.2. Material	39
10.2.3. Listening Laboratory	40
10.2.4. Degradation Factors	40
10.3. Application Developers	41
10.4. Codec Implementers	42
10.5. End Users	42
11. Security Considerations	42
12. IANA Considerations.....	42
13. References	43
13.1. Normative References	43
13.2. Informative References	43
14. Acknowledgments	48

1. Introduction

The IETF Working Group CODEC is standardizing an Internet Interactive Audio and Speech Codec (IIAC). If the codec shall be of high quality it is important to measure the codec's quality throughout the entire process of development, standardization, and usage. Thus, this document supports the standardizing process by providing an overview of quality metrics, quality assessment procedures, and other quality control issues and gives suggestions on how to test the IIAC.

Quality must be measured by the following stakeholders and in the following phases of the codec's development:

- o Codec developers must decide on different algorithms or parameter sets during the development and enhancement of a codec. These might also include the selection among multiple codec candidates that implement different algorithms; however the WG Codec base its work on a common consensus not on a competitive selection of one of multiple codec contributions. Thus, measuring the quality of codecs to select one might not be required. Besides selection, one is obliged to debug the codec software. To find errors and bugs - and programming mistakes are present in any complex software - the developer has to test this software by conducting quality measurements.
- o Typically the codec standardization includes a qualification phase that measures the performance of a codec and verifies whether it confirms to predefined quality requirements. In the qualification phase, it becomes obvious whether the codec development and standardization has been successful. Again, in the process of rigorous testing during qualification phase, algorithmic weaknesses and bugs in the implementation may be found. Still, in complex software such as the IIAC, correctness cannot be proved or guaranteed.

- o Users of the codec need to know how well the codec is performing while manufactures need to decide whether to include the IIAC in their products. Quality measures play an important role in this decision process. Also, the numerous quality measurement results of the quality help developers of the VoIP system to dimension or tune their system to take optimal advantage of a codec. For example, during network planning, operators can predict the amount of bandwidth needed for high quality voice calls. An adaptive VoIP application needs to know which quality is achieved with a different codec parameters set to be able to make an optimal selection of the codec parameters under varying network conditions.
As suggested in [50] an RTP payload specification for an IIAC codec should include a rate control. Similar to the performance of the codec, the rate control unit has a big impact on the overall quality of experience. Thus, it should be tested well too.
- o Software implementers need to verify whether their particular codec implementation that might be optimized on a specific platform confirms to the standard's reference implementation. This is particularly important as some intellectual property rights might only be granted, if the codec conforms to the standard.
As the IIAC must not to be bit conform, which would allow simple comparisons of correctness, other means of conformance testing must be applied.
In addition, the standard conformance and interoperability of multiple implementations must be checked.
Last but not least, implementers may implement optimized concealment algorithms, jitter buffers or other algorithms. Those algorithms have to be tested, too.
- o Since the success of MP3, end users do acknowledge the existence of a high quality codec. It would make sense to use the IIAC in a brand marketing campaign (such as "Intel inside"). A quality comparison between IIAC and other codecs might be part of the marketing. Online testing with user participation might also raise the awareness level.

All those stakeholders might have different requirements regarding the codec's quality testing procedures. Thus, this document tries to identify those requirements and shows which of the existing quality measurement procedures can be applied to fulfill those specific demands efficiently.

In the following section we describe a primary optimization goal: Quality of Experience (QoE). Next, we briefly list the most common methods of how to perform subjective evaluations on speech and audio quality. In Section 4, 5, and 6, we discuss on how to measure complexity, latency, and bit- and frame rates. Section 7 describes how other SDOs have measured the quality of their codecs. As compared IIAC to previous standardized codecs, the IIAC is likely to have different unique requirements and thus needs newly developed quality testing procedures. To achieve this, in Section 8 we describe the properties of Internet transmission paths. Section 9 summarizes the usage scenarios, for which the codec is going to be used and finally, in Section 10, we recommend procedures on how to test the IIAC.

2. Optimization Goal

The aim of the Codec WG is to produce a codec of high quality. However, how can quality be measured? The measurement of the features of a codec can be based on many different criteria. Those include complexity, memory consumption, audio quality, speech quality, and others. But in the end, it's the users' opinions that really count since they are the customers. Thus, one important - if not the most important quality measure of the IIAC - shall be the Quality of Experience (QoE).

The ITU-T Standards ITU-T P.10/G.100 [22] defines the term "Quality of Experience" as "the overall acceptability of an application or service, as perceived subjectively by the end-user." The ITU-T document G.RQAM [21] extends this definition by noting that "quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.)" and that the "overall acceptability may be influenced by user expectations and context".

These definitions already give guidelines on how to judge the quality of the IIAC:

- o The acceptability and the subjective quality impression of endusers have to be measured (Section 3).
- o The IIAC codec has to be tested as part of an entire telecommunication system. It must be carefully considered whether to measure the codec's performance just in a stand-alone setup or to evaluate it as part of the overall system (Section 8).

- o The environments and contexts of particular communication scenarios have to be considered and controlled because they have an impact on the human rating behavior and on quality expectations and requirements (Section 9).

3. Measuring Speech and Audio Quality

The perceived quality of a service can be measured by various means. If humans are interrogated, those quality tests are called subjective. If the tests are conducted by instrumental means (such as an algorithm) they are called objective. Subjective tests are divided up into formal and informal tests. Formal tests follow strictly defined procedures and methods and typically include a large number of subjects. Informal tests are less precise because they are conducted in an uncontrolled manner.

3.1. Formal Subjective Tests

Formal subjective tests must follow a well-defined procedure. Otherwise the results of multiple tests cannot be mutually compared and are not repeatable. Most subjective testing procedures have been standardized by the ITU. If applied to coding testing, the testing procedures follow the same pattern [26]:

"Performing subjective evaluations of digital codecs proceeds via a number of steps:

- o Preparation of source speech materials, including recording of talkers;
- o Selection of experimental parameters to exercise the features of the codec that are of interest;
- o Design of the experiment;
- o Selection of a test procedure and conduct of the experiment;
- o Analysis of results."

The ITU has standardized different formal subjective tests to measure the quality of speech and audio transmission, which are described in the following.

3.1.1. ITU-R Recommendation BS.1116-1

The ITU-R BS.1116-1 standard [14] is good for audio items with small degradations (stimuli) and uses a continuous scale from

imperceptible (5.0) to very annoying (1.0). It is a double blind triple-stimulus with a hidden reference testing method and must be done twice for the degraded sample and the hidden reference. In a 30 minutes session, 10-15 sample items can be judged. Overall, about 20 subjects shall rate the items. Testing shall take place with loudspeakers in a controlled environment or with headphones in a quiet room.

3.1.2. ITU-R Recommendation BS.1534-1 (MUSHRA)

The ITU-R BS.1534-1 standard [16] defines a method for the subjective assessment of intermediate quality levels. Multiple audio stimuli are compared at the same time. Maximal 12 but preferably only 8 stimuli plus a hidden one with Hidden Reference and an anchor are compared and judged. MUSHRA uses a continuous quality scale (CQS) ranging from 0 to 100 divided into five equal intervals ("bad" to "excellent"). In 30 minutes, about 42 stimuli can be tested. Again, 20 test subjects shall rate the items with either headphones or loudspeakers.

The standard recommends using as lower anchor a low-pass filtered version with a bandwidth limit of 3.5 kHz. Additional anchors are recommended, especially if specific distortions are to be tested.

3.1.3. ITU-T Recommendation P.800

The ITU-T P.800 defines multiple testing procedures to assess the speech quality of telephone connections. The most important procedure is called listening-only speech quality of telephone connections. Listeners rate short groups of unrelated sentences. The listeners are taken from the normal telephone-using population (no experts). They use a typical sending system (e.g. a local telephone) that may follow "modified IRS" frequency characteristics. The results is the listening-quality scale, which is an absolute category scale (ACS) ranging from excellent=5 to bad=1. Listeners can judge about 54 stimuli within 30 minutes.

Other tests described in P.800 measure listening-effort, loudness-preference scale, conversation opinion and difficulty, delectability, degradation, or minimal differences.

3.1.4. ITU-T Recommendation P.805

The P.805 standard [24] extends P.800 and defines precisely how to measure conversational quality. Subjects have to do conversation tests to evaluate the communication quality of a connected. Expert, experienced or untrained (naive) subjects have to do these tests

collaboratively in soundproof cabinets. Typically, 6 transmission conditions can be tested within 30 minutes. Depending on the required precision, these tests have to be made 20 to 40 times.

3.1.5. ITU-T Recommendation P.880

To measure time-variable distortion, a continuous evaluation of speech quality has been defined in P.880 [31]. Subjects have to assess transmitted speech quality consisting of long speech sequences with quality/time fluctuations. The quality is rated on a continuous scale ranging from Excellent=5 to Bad=1 is dynamically changed over the time while the stimuli are played. Stimuli have a length of between 45 seconds and 3 minutes.

3.1.6. Formal Methods Used for Codec Testing at the ITU

In the last year, new narrow and wideband codecs have been tested using ITU-T P.800 (and ITU-T P.830). For the ITU-T G.719 standard, which supports besides speech content also audio, the ITU-R BS.1116-1 testing method has been applied during the selection of potential codec candidates. During the qualification phase, the method that was used was the ITU-P BS.1584-1. For the ITU-T G.718 codec, the Absolute Category Rating (ACR) following ITU-T P.800 has been applied.

3.2. Informal Subjective Tests

Besides formal tests, informal subjective tests following less stringent conditions might be taken to judge the quality of stimuli. However, informal tests cannot be easily verified and lack the reliability, accuracy and precision of formal tests. Informal tests are needed if the available number of subjects who are able to conduct the tests is low, or if time or money is limited.

3.3. Interview and Survey Tests

In ITU-T P.800 [23] and [9] interview and survey tests are described. In P.800, it says that "if the rather large amount of effort needed is available and the importance of the study warrants it, transmission quality can be determined by 'service observations'."

These service observations are based on statistical surveys common in social science and marketing research. Typically, the questions asked in a survey are structured.

In addition, according to [23]: "To maintain a high degree of precision a total of at least 100 interviews per condition is required. A disadvantage of the service-observation method for many purposes is that little control is possible over the detailed characteristics of the telephone connections being tested."

3.4. Web-based Testing

If the large-wide scale proliferation of the Internet, researchers suggested testing the speech or audio quality on web sites via web site visitors [43]. A current web site that compares multiple audio codecs has been setup at SoundExpert.org [42]. On this web site, a user can download an audio item that consists of a reference item and a degraded item. Then, the user must identify the reference and rate the ODG of the degraded item. The tests are single-blind as the user does not know which codec he is currently rating.

One can anticipate that the visitors of web sites will use similar equipment for testing of audio samples and for conducting VoIP calls. Thus, web site testing can be made realistic in a way that considers the impact of (typically used) loudspeakers and headphones.

However, currently used web sites lack a proper identification of outliers. Thus, all ratings of all users are considered despite the fact that they might be (deliberately) faked or that subjects might not be able to hear well the acoustic difference. Thus, one can expect that web based ratings will show a high degree of variation and that many more tests are needed to achieve the same confidence that is gained within formal tests. A profound scientific study on the quality of web based audio rating has not yet been published. Thus, any statements on the validity of web based rating are premature.

3.5. Call Length and Conversational Quality

In the ETSI technical report document ETR-250 [6], a model is presented that discusses various impairments caused in narrow band telephone systems. The ETSI model describes the combinatorial effect of all those impairments. The ETSI model later became the famous E-Model described in ITU-T G.107. Both the ETSI- and the E-Model calculate the R factor that ranges from 0 (bad) to 100 (excellent conversational quality).

Based on the R factor, the users' reaction to the voice transmission quality of a connection can be predicted. For example, Section 8.3 describes the effect that users terminate the call if the quality is

bad. More precisely, they summarize it as users who "(i) terminate their calls unusually early, (ii) re-dial or even (iii) actually complain to the network operator".

In the ETSI model, the percentage of users "terminating calls early", TME, is given as

$$TME = 100 \cdot \operatorname{erf}\left(\frac{36-R}{16}\right) \%$$

with $\operatorname{erf}(X)$ being the sigmoid shaped Gaussian error function and R the R Factor of the E-Model (Figure 1). This relation is based on results from "AT&T Long toll" interviews as cited in [2].

These findings have been confirmed by Holub et al. [12] who have studied the correlation between call length and narrow band speech quality. Birke et al. [1] have also studied the duration of phone calls which show a duration varying with day time and day of the week and also may be affected by pricing schemata.

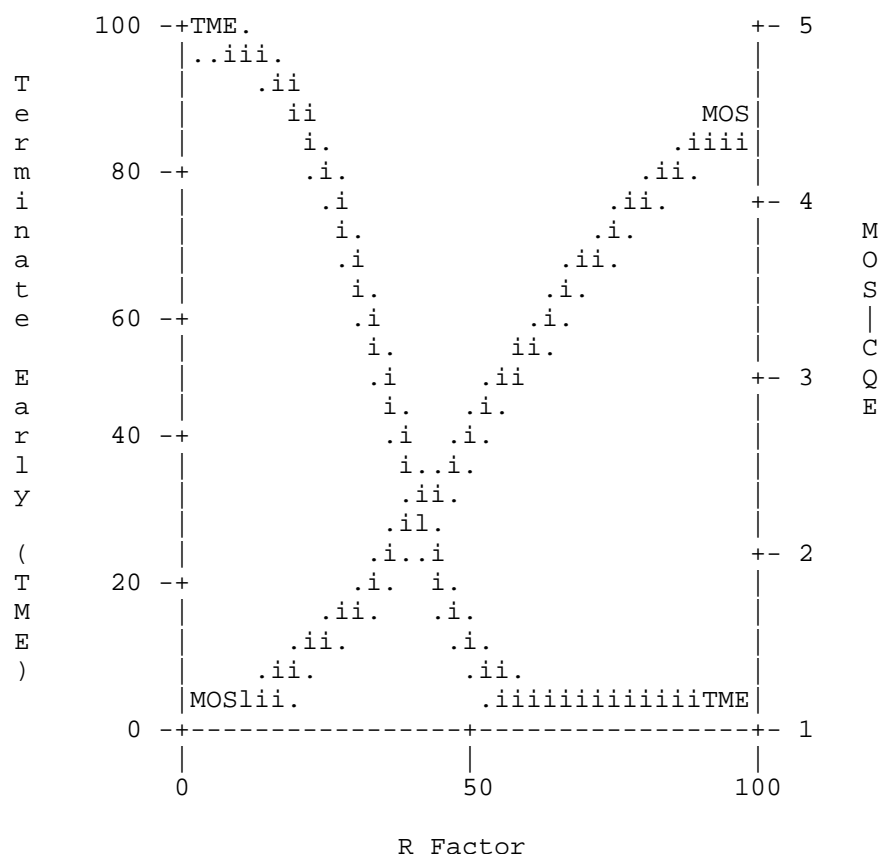


Figure 1 - Relation between calls terminating early, the R Factor, and the speech quality given in (MOS-CQE)

Whereas bad quality is related to short calls, it remains unproven whether better quality (>4 MOS) results in longer phone calls. There are two factors which might have an opposite effect on the call length. On the one hand, if the quality is superb, the talkers might be more willing to talk because of the pleasure of talking, on the other hand they might fulfill their conversational tasks faster because of the great quality. Thus, depending on the context, good speech quality might result either in longer or shorter calls.

3.6. Field Studies

Field studies can be conducted if usage data on calls are collected. Field studies are useful to monitor real user behavior and to collect data about the actual conversational context.

Because of highly varying conditions, the precision of those measurements is high and many tests have to be done to get significantly different measurement values. Also, the tests are not repeatable because the conditions are changing with time.

For example, Skype has done quality tests in a deployed VoIP system in the field with its users as testers [47]. The subjective tests are done in the following manner.

- o Download of test vectors to VoIP clients. Typically, this can be done with an automated software update.
- o Delivery changing VoIP configurations (such as the used codecs) so that different calls are subjected to different configurations. The selection of configurations can be done randomly, alternating in time or based on other criteria.
- o Collecting feedback from the users. For example, the following parameters can be monitored or recorded:
 - o The call length and other call specific parameters
 - o A user's quality voting (e.g. MOS-ACR) after the call
 - o Other feedback of the user (e.g. via support channels)

The field tests have the benefit of being conducted under real conditions with the real users. However, they have some drawbacks. First, the experimental conditions cannot be controlled well. Second, the tests are only valid for the current situations and do not allow predictions for other use cases. Third, the statistical significance might be largely questionable if confidence intervals are overlapping.

The costs for running the tests are low because the users are doing the tests for free. However, the operator might lose users after a user experienced a test case causing bad quality.

3.7. Objective Tests

Objective tests, also called instrumental tests, try to predict the human rating behavior with mathematical models and algorithms. They also calculate quality ratings for a given set of audio items. Naturally, they are not rating as precisely as their human counterparts, whom they try to simulate. However, the results are repeatable and less costly than formal subjective testing campaigns. Instrumental methods have a limited precision. That means that their

quality ratings do not perfectly match the results of formal listening-only tests. Typically, the correlation between formal results and instrumental calculations are compared using a correlation function. The resulting metric is given as R ranging from 0 (no correlation) to 1 (perfect match).

Over the last years, several objective evaluation algorithms have been developed and standardized. We describe them briefly in the following.

3.7.1. ITU-R Recommendation BS.1387-1

The ITU developed an algorithm that is called Perceptual Evaluation of Audio Quality (PEAQ). It was published in the document ITU-R BS.1387 called Method for objective measurements of perceived audio quality in 1998 [15]. PEAQ is intended to predict the quality rating of low-bit-rate coded audio signals. Two different versions of PEAQ are provided: a basic version with lower computational complexity and an advanced version with higher computational complexity.

PEAQ calculates a quality grading called "Objective Difference Grade" (ODG) ranging from 0 to -4. Typically, it shows a prediction quality of between $R=0.85$ and 0.97 when compared to subjective testing results. The ITU-T Study Group 12 assumes that PEAQ can detect auditable differences between two implementations of the same codec [5].

3.7.2. ITU-T Recommendation P.862

The ITU-T PESQ algorithm [27] is intended to judge distortions caused by narrow band speech codecs and other kind of channel and transmission errors. These include also variable delays, filtering and short localize distortions such as those caused by frame loss concealment. For a large number of conditions, the validity and precision of PESQ has been proven. For untested distortions, prior subjective tests must be conducted to verify whether PESQ judges these kinds of distortions precisely. Also, it is recommended to use PESQ for 3.1 kHz (narrow-band) handset telephony and narrow-band speech codecs only. For wide-band operations, a modified filter has to be applied prior to the tests.

Furthermore, the ITU-T Recommendation P.862.1 [28] describes how to transfer the PESQ's raw scores, which range from -0.5 to 4.5 , to MOS-LQO values similar to those gathered from ACR ratings. Then, as it has been shown, the correlation between a large corpus of testing samplings shows a correlation of $R=0.879$ (instead of $R=0.876$) between subjective and MOS-LQO (respective PESQ raw) ratings. The

ITU-T Recommendation P.862.2 [29] modifies the PESQ algorithm slightly to support wideband operations. And finally, the ITU-T Recommendation P.862.3 [30] gives detailed hints and recommendations on how and when to use the PESQ algorithms.

3.7.3. ITU-T Draft P.OLQA

The soon-to-be standardized algorithm P.OLQA [40] extends PESQ and will be able to rate narrow to super-wideband speech and the effect of time-varying speech playout. Later distortions are common in modern VoIP systems which stretch and shrink the speech playout during voice activity to adapt it to the delay process of the network.

4. Measuring Complexity

Besides audio and speech quality, the complexity of a codec is of prime importance. Knowing the algorithmic efficiency is important because:

- . the complexity has an impact on power consumption and system costs
- . the hardware can be selected to fit pre-known complexity requirements and
- . different codec proposals can be compared if they show similar performances in other aspects.

Before any complexity comparisons can be made, one has to agree on an objective, precise, reliable, and repeatable metric on how to measure the algorithmic efficiency. In the following, we list three different approaches.

4.1. ITU-T Approaches to Measuring Algorithmic Efficiency

Over the last 17 years, the ITU-T Study Group 16 measured the complexity of codecs using a library called ITU-T Basic Operators and described in ITU-T G.191 [19], which counts the kind and number of operations and the amount of memory used. The latest version of the standard supports both fix-point operations of different widths and floating operations. Each operation can be counted automatically and weighted accordingly. The following source code is an [edited] excerpt from the source file baseop32.h:

```

/* Prototypes for basic arithmetic operators */

/* Short add,          1 */
Word16 add (Word16 var1, Word16 var2);

/* Short sub,          1 */
Word16 sub (Word16 var1, Word16 var2);

/* Short abs,          1 */
Word16 abs_s (Word16 var1);

/* Short shift left,   1 */
Word16 shl (Word16 var1, Word16 var2);

/* Short shift right,  1 */
Word16 shr (Word16 var1, Word16 var2);

...

/* Short division,     18 */
Word16 div_s (Word16 var1, Word16 var2);

/* Long norm,          1 */
Word16 norm_l (Word32 L_var1);

```

In the upcoming ITU-T G.GSAD standard another approach has been used as shown in the following code example. For each operation, WMPOS functions have been added, which count the number of operations. If the efficiency of an algorithm has to be measured, the program is started and the operations are counted for a known input length.

```

for (i=0; i<NUM_BAND; i++)
{
#ifdef WMOPS_FX
    move32();move32();
    move32();move32();
#endif
    state_fx->band_enrg_long_fx[i] = 30;
    state_fx->band_enrg_fx[i] = 30;
    state_fx->band_enrg_bgd_fx[i] = 30;
    state_fx->min_band_enrg_fx[i] = 30;
}

```


4.2. Software Profiling

The previously described methods are well-established procedures on how to measure computational complexity. Still, they have some drawbacks:

- o Existing algorithms must be modified manually to include instructions that count arithmetic operations. In complex codecs, this may take substantial time.
- o The CPU model is simple as it does not consider memory access (e.g. cache), parallel executions, or other kinds of optimization that are done in modern microprocessors and compilers. Thus, the number of instructions might not correlate to the actual execution time on modern CPUs.

Thus, instead of counting instructions manually, run times of the codec can be measured on a real system. In software engineering, this is called profiling. The Wikipedia article on profiling [54] explains profiling as follows:

"In software engineering, program profiling, software profiling or simply profiling, a form of dynamic program analysis (as opposed to static code analysis), is the investigation of a program's behavior using information gathered as the program executes. The usual purpose of this analysis is to determine which sections of a program to optimize - to increase its overall speed, decrease its memory requirement or sometimes both.

- o A (code) profiler is a performance analysis tool that, most commonly, measures only the frequency and duration of function calls, but there are other specific types of profilers (e.g. memory profilers) in addition to more comprehensive profilers, capable of gathering extensive performance data
- o An instruction set simulator which is also - by necessity - a profiler, can measure the totality of a program's behaviour from invocation to termination."

Thus, a typical profiler such as the GNU gprof can be used to measure and understand the complexity of a codec implementation. This is precisely the case because it is used on modern computers. However, the execution times depend on the CPU architecture, the PC in general, the OS and parallel running programs.

To ensure repeatable results, the execution environment (i.e. the computer) must be standardized. Otherwise the results of run times cannot be verified by other parties as the results may differ if done under slightly changed conditions.

4.3. Cycle Accurate Simulation

If reliable and repeatable results are needed, another similar approach can be chosen. Instead of run times, CPU clock cycles on a virtual reference system can be measured. Quoting Wikipedia again [52]:

"A Cycle Accurate Simulator (CAS) is a computer program that simulates a microarchitecture cycle-accurate. In contrast an instruction set simulator simulates an Instruction Set Architecture usually faster but not cycle-accurate to a specific implementation of this architecture."

With a cycle accurate simulator, the execution times are precise and repeatable for the system that is being studied. If two parties make measurements using different real computers, they still get the same results if they use the same CAS.

A cycle accurate simulator is slower than the real CPU by a factor of about 100. Also, it might have a measurement error as compared to the simulated, real CPU because the CPU is typically not perfectly modeled.

If an x86-64 architecture shall be simulated, the open-source Cycle accurate simulator called PTLsim can be considered [55]. PTLsim simulates a Pentium IV. On their website, the authors of PTLsim write:

"PTLsim is a cycle accurate x86 microprocessor simulator and virtual machine for the x86 and x86-64 instruction sets. PTLsim models a modern superscalar out of order x86-64 compatible processor core at a configurable level of detail ranging from full-speed native execution on the host CPU all the way down to RTL level models of all key pipeline structures."

Another cycle accurate simulator called FaCSIM simulated the ARM9E-S processor core and ARM926EJ-S memory subsystem [36]. It is also available as open-source. Texas Instruments also provides a CAS for its C64x+ digital signal processor [44].

To have a metric that is independent of a particular architecture, the results of cycle accurate simulators could be combined.

4.4. Typical run time environments

The IIAC codec will run on various different platforms with quite diverse properties. After discussions on the WG mailing list, a few typical run time environments have been identified.

Three of the run time environments are end devices (aka phones). The first one is a PC, either stationary or a portable, having a >2 GHz PCU, >2 GByte of RAM, and a hard disk for permanent storage. Typically, a Windows, MacOS or Linux operating system is running on a PC. The second one is a SmartPhone, for example with an ARM11 500 MHz CPU, 192 Mbyte RAM and 256 MByte Flashrom. An example is the HTC Dream Smart phone equipped with Qualcomm MSM7201A chip. Various operating systems are found on those devices such as Symbian, Android, and iOS. The last ones are high end stationary VoIP phones with for example a 275-MHz MIPS32 CPU (with 400 DMIPS) with a 125-MHz (250 MIPS) ZSP DSP with dual-MAC. They both have more than 1 Mbyte RAM and FlashRom. An exemplary Chip is the BCM1103 [3].

Besides phones, VoIP gateways are frequently needed for conferencing or transcoding to legacy VoIP or PSTN. In this case, two different platforms have been identified. The first one is based on standard PC server platforms. It consists, for example, of an Intel six core Xeon 54XX or 55XX, two 1 GB NIC, 12 GByte RAM, hard disks, and a Linux operating system. Thus, a server can serve from 400 to 10000 calls depending on conference mode, codecs used, and ability of user pre-encoded audio [46]. On the other hand, high density, highly optimized voice gateways use a special purpose hardware platform like for example, TNETV3020 chips consisting of six TI C64x+ DSPs with 5.5 MB internal RAM. If they run with a Telogy conference engine, they might serve about 1300 AMR or 3000 G.711 calls per chip [45].

5. Measuring Latency

Latency is a measure of time delay experienced in a system. Latency can be measured as one-way delay or as round-trip time. The latter one is the one-way latency from a source to destination plus the one-way latency back from destination to source. Latency can be measured at multiple positions, at the network layer or at higher layers [53].

As we aim to increase the Quality of Experience, the mouth-to-ear delay is of importance because it directly correlates with perceptual quality [17]. More precisely, the acoustic round-trip time shall be a means of optimization when studying interactive and conversational application scenarios.

5.1. ITU-T Recommendation G.114

The G.114 standard [45] gives guidelines on how to estimate one-way transmission delays. It describes how the delay introduced by the codec is generated. Because most of the encoders do a processing of frames, the duration of a frame (named "frame size") is the foremost contributor to the overall algorithmic delay. Citing [18]:

"In addition, many coders also look into the succeeding frame to improve compression efficiency. The length of this advance look is known as the look-ahead time of the coder. The time required to process an input frame is assumed to be the same as the frame length since efficient use of processor resources will be accomplished when an encoder/decoder pair (or multiple encoder/decoder pairs operating in parallel on multiple input streams) fully uses the available processing power (evenly distributed in the time domain). Thus, the delay through an encoder/decoder pair is normally assumed to be:"

$2 \times \text{frameSize} + \text{lookAhead}$

In addition, if the link speeds are low, the serialization delay might contribute significantly to the codec delay.

Also, if IP transmissions are used and multiple frames are concatenated in one IP packet, further delay is added. Then, "the minimum delay attributable to codec-related processing in IP-based systems with multiple frames per packet is:"

$(N+1) \times \text{frameSize} + \text{lookAhead}$

"where N is the number of frames in each packet."

5.2. Discussion

Extensive discussion on the WG mailing list led to the insight that the afore mentioned ITU delay model overestimates the delay introduced by the codec. In the last decade, two developments led to slightly other conditions.

First, the processing power of CPU increased significantly (see Section 4.4). Nowadays, even stand-alone VoIPs have CPUs with a speed of 300 MHz. They are capable of doing the encoding and decoding faster than real time. Thus, also the delay introduced by processing is not at 100% anymore but significantly lower. For example, it might be just 10% or less.

Second, even if the CPUs are fully loaded, especially if also other tasks such as a video conference or other calls need to be processed, advantaged scheduling algorithms allow for a timely encoding and decoding. For example, a staggered processing schedule can be used to reduce processing delays [45].

Thus, the impact of processing delay is reduced significantly in most of the cases.

Moreover, besides a look-ahead time, the decoder might also contribute to the algorithmic delay e.g. if decoded and concealed periods shall be mixed well.

6. Measuring Bit and Frame Rates

For decades, there was a quest to achieve high quality while keeping the coding rate low. Coding rate, sometimes called multimedia bit rate, is the bit rate that an encoder produces as its output stream. In cases of variable rate encoding, the coding bit rate differs over time. Thus, one has to describe the coding rate statistically. For example, minimal, mean, and maximal coding rates need to be measured.

A second parameter is the frame rate as the encoder produces frames at a given rate. Again, in case of discontinuous transmission modes (DTX), the frame rate can vary and a statistical description is required.

Both coding and frame rate influence network related bit rates. For example, the physical layer gross bit rate is the total number of physically transferred bits per second over a communication link, including useful data as well as protocol overhead [51]. It depends on the access technology, the packet rate, and packet sizes. The physical layer net bit rate is measured in a similar way but excludes the physical layer protocol overhead. The network throughput is the maximal throughput of a communication link of an access network. Finally, the goodput or data transfer rate refers to the net bit rate delivered to an application excluding all protocol headers and data link layer retransmissions, etc. Typically, to avoid packet losses or queuing delay, the goodput shall be equally large as the coding rate.

The relation between goodput and the physical layer gross bit rate is not trivial. First of all, the goodput is measured end-to-end. The end-to-end path can consist of multiple physical links, each having a different overhead. Second, the overhead of physical layers may vary with time and load, depending for example on link

utilization and link quality. Third, packets may be tunneled through the network and additional headers (such as IPsec) might be added. Fourth, IP header compression might be applied (as in LTE networks) and the overhead might be reduced. Overall, many information about the network connection must be collected to predict what the relation between physical layer gross bit rate and a given coding and frame rate is going to be. Applications, which have only a limited view of the network, can hardly know the precise relation.

For example, the DCCP TFRC-SP transport protocol simply estimates a header size on data packets of 36 bytes (20 bytes for the IPv4 header and 16 bytes for the DCCP-Data header with 48-bit sequence numbers) [7][8]. Thus, [11] suggested a typical scenario in which one encoded frame is transmitted with the RTP, UDP, IPv4 and IEEE 802.3 protocols and thus each packet contains packet headers having 12 bytes, 8 bytes, 20 bytes and 18 bytes respectively. The gross bit rate calculates as

$$r_{\text{gross}} = r_{\text{coding}} + \text{overhead} \cdot \text{framerate}$$

where r_{coding} is the coding rate of the encoding, framerate is the frame rate of the codec, overhead is the number of bits for protocol headers in each packet (typically $58 \cdot 8 = 464$), and the r_{gross} is the rate used on physical mediums.

7. Codec Testing Procedures Used by Other SDOs

To ensure quality, each newly standardized codec is rigorously tested. ITU-T Study Group 12 and 16 have developed very good and mature procedures on how to test codecs. The ITU-T Study Group 12 has described the testing procedures of narrow- and wide-band codecs in the ITU-T P.830 standard.

7.1. ITU-T Recommendation P.830

The ITU-T P.830 recommendation describes methods and procedures for conducting subjective performance evaluations of digital speech codecs. It recommends for most applications the Absolute Category Rating (ACR) method using the Listening Quality scale. The process of judging the quality of a speech codec consists of five steps, which are described in the following.

Step 1: Preparation of Source Speech Materials Including Recording of Talkers. When testing a narrow band codec, the recommendation suggests to use a bandwidth filter before applying sample items to a codec. This bandwidth filter is called modified Intermediate Reference System (IRS) and limits the frequency band to the range

between 300 and 3400 Hz. In addition, the recommendation states that "if a wideband system (100-7000 Hz) is to be used for audio-conferencing, then the sending end should conform to IEC Publication 581.7."

It also says that "speech material should consist of simple, short, meaningful sentences." The sentences shall be understandable to a broad audience and sample items should consist of two or three sentences, each of them having a duration of between 2 and 3 seconds. Sample items should not contain noise or reverberations longer than 500 ms. The recommendation also makes suggestions on the loudness of the signal: "A typical nominal value for mean active speech level (measured according to Recommendation P.56) is -20 dBm0, corresponding to approximately -26 dBov"

Step 2: Selection of Experimental Parameters to Exercise the Features of the Codec That Are of Interest. Various parameters shall be tested. Those include

- o Codec Conditions

- o Speech input levels ("input levels of 14, 26 and 38 dB below the overload point of the codec")
- o Listening levels ("levels should lie 10 dB to either side of the preferred listening level")
- o Talkers
 - . Different talkers ("a minimum of two male and two female talkers")
 - . Multiple talkers ("multiple simultaneous voice input signals")
- o Errors ("randomly distributed bit errors" or burst-errors)
- o Bitrates ("The codec must be tested at all the bit rates")
- o Transcodings ("Asynchronous tandeming", "Synchronous tandeming", and "Interoperability with other speech coding standards")
- o Mismatch (sender and receiver operate in different modes)
- o Environmental noise (sending) ("30 dB for room noise" and "10 dB and 20 dB for vehicular noise")

- o Network information signals ("signaling tones, conforming to Recommendation Q.35, should be tested subjectively, and the minimum should be proceed to dial tone, called subscriber ringing tone, called subscriber engaged tone, equipment engaged tone, [and] number unobtainable tone.")
- o Music ("to ensure that the music is of reasonable quality")
- o Reference conditions ("for making meaningful comparisons")
 - o Direct (no coding, only input and output filtering)
 - o Modulated Noise Reference Unit (MNRU)
 - o Signal-to-Noise Ratio (SNR) (for comparison purposes)
 - o Reference codecs

Step 3: Design of the Experiment. The considerations described in B.3/P.80 apply here. Typically, it is not possible to test each combination of parameters. Thus, recommendation P.830 states that "it is recommended that a minimum set of experiments be conducted, which, although they would not cover every combination, would result in sufficient data to make sensible decisions. [...] Extreme caution should be used when comparing systems with widely differing degradations, e.g. digital codecs, frequency division multiplex systems, vocoders, etc., even within the same test."

Step 4: Selection of a Test Procedure and Conduct of the Experiment. Here, the considerations as in B.4/P.80 apply. However, a modified IRS at the receiver shall be used (narrow band) or an IEC Publication 581.7 filter (wideband). Also, "Gaussian noise equivalent to -68 dBmp should be added at the input to the receiving system to reduce noise contrast effects at the onset of speech utterances."

Step 5: Analysis of Results. Again, the considerations detailed in B.4.7/P.80 apply. The arithmetic mean (over subjects) is to be calculated for each condition at each listening level.

7.2. Testing procedure for the ITU-T G.719

Recently, the ITU-T has standardized the audio and speech codec ITU-T G.719. The G.719 has similar properties as the anticipated IIAC, thus the optimization and characterization of the G.719 is of particular interest.

In the following, we will describe the "Quality Assessment Test Plan" in TD 322 and 323 [33][35]. The ITU Study Group 16 used ITU-R BS.1116 to tests sample items. Audio sample items were sampled at 48 kHz mixed down to mono. Speech sample items contain one sentence with a duration of 4 s, mixed content had a duration of 5-6 s and music a duration of between 10 and 15 s. The beginning and ending of the samples were smoothed. Also, a filter was applied to limit the nominal bandwidth of the input signal to the range of 20 to 20000 Hz. As for the mixed content, advertisements, film trailers and news (including a jingle) have been selected. For music items, classical and modern styles of music have been selected. Besides the codec under test, test stimuli degraded with LAMP MP3 and G722 were added to the tests. Some test stimuli have been modified to include reverberations or an interfering talker and office noise. Some tests were done studying the effect of a frame erasure rate of 3% having random loss patterns. All listening labs used different sample items and attention paid to not use the same material twice.

Listening labs were required to provide the results of 24 experienced listeners excluding those listeners, who did not passed a pre- and post-screening. The experienced listeners should "neither have a background in technical implementations of the equipment under test nor do they have detailed knowledge of the influence of these implementations on subjective quality".

During the tests, "circum aural headphones - open back for example: STAX Signature SR-404 or Sennheiser HD-600) on both ears (diotic presentation)" were used. The listening levels were -26 dB relative to OVL.

Some results of the listening tests are given in TD 341 R1 [34]. In those tests, they also compared the subjective ratings that were made following BS.1116 with the objective ratings of ITU-R BS.1387-1. The correlation between objective and subjective ratings was below $R=0.9$.

8. Transmission Channel

Between speech encoder and decoder lies a transmission channel that effects the transmission. For cellular or wireless phones, the typical transmission channel is assumed to be equal to the wireless link(s). This typically means, that a circuit switch link is assumed (e.g., in GSM, UMTS, DECT). The bandwidth is typically constant in DECT and GSM or variable in a given range depending on the quality of the wireless transmission (UMTS). Bit errors do occur but they don't be equally distributed if unequal bit error correction is applied (UMTS).

In the case of the IIAC codec, the transmission channel is the internet. More precisely, it is the packet transmission over the Internet, plus the transport protocol (e.g. UDP, TCP, DCCP), plus potentially Forward Error Correction, and plus dejittering buffers.

Also, the transmission channel is reactive. It changes its properties depending on how much data is transmitted. For example, parallel TCP flows reduce their transmission bandwidth in the presence of an unresponsive UDP stream.

Overall, one can say that the transmission channel "Internet" is difficult to understand. Thus, in this chapter, we try to shed light on the question of what types of transmission channels a codec has to cope with.

8.1. ITU-T G.1050: Network Model for Evaluating Multimedia Transmission Performance over IP (11/2007)

The current ITU-T G.1050 standard [20] describes layer 3 packet transmission models that can be used to evaluate IP applications. The models are of statistical nature. They consider networks architectures, types of access links, QoS controlled edge routing, MTU size, networks faults, link failures, route flapping, reordered packets, packet loss, one-way delay, variable deploys and background traffics.

G.1050 is a network model consisting of three parts, LAN a, LAN b, and an interconnection core. Both LANs can have different rates and occupancy and can be of different types. LAN and core are connected via access technologies, which might vary in data rate, occupancy and MTU size.

The core is characterized by route flapping, link failures, one-way delay, jitter, packet loss and reordered packets. Route flaps are repeatedly changed in a transmission path because of alternating routing tables. These routing updates cause incremental changes in the transmission delays. A link failure is a period of consecutive packet loss. Packet losses can be bursty having a high loss rate during bursts and having otherwise a lower loss rate otherwise. Delays are modeled via multiple different jitter models supporting delay spikes, random jitter and filtered random jitters.

The standard recommends three profiles, named "Well-managed IP network", "Partially-managed IP network", and "Unmanaged IP Network, Internet", which differ in their connection qualities.

Limitations to these models are the missing cross-correlation between packet delays and packet loss events, the lack of responsiveness to the tests application flow, and the lack of link qualities that vary with time.

8.2. Draft G.1050 / TIA-921B

Currently, an enhancement to ITU-T G.1050 (11/2007) is being developed (e.g. [13])). It does not use a statistical model but takes advantage of the NS/2 simulator. Thus, most of the above mentioned limitations have been overcome.

Despite that, even the new model does not yet give an answer to the question of which distributions of typical Internet connection qualities can be expected.

8.3. Delay and Throughput Distributions on the Global Internet

In general, it is not precisely known how the qualities of end-to-end connections are distributed. It is also unclear whether the anticipated IIAC Codec will be used globally or whether its area of usage will be somehow restricted.

Despite the fact, that the codec has to be optimized for an unknown Internet, the following scientific publications give an estimate on how different Internet end-to-end paths might behave. One recent example is on studies about the residential broadband Internet access traffic of a major European ISP [37].

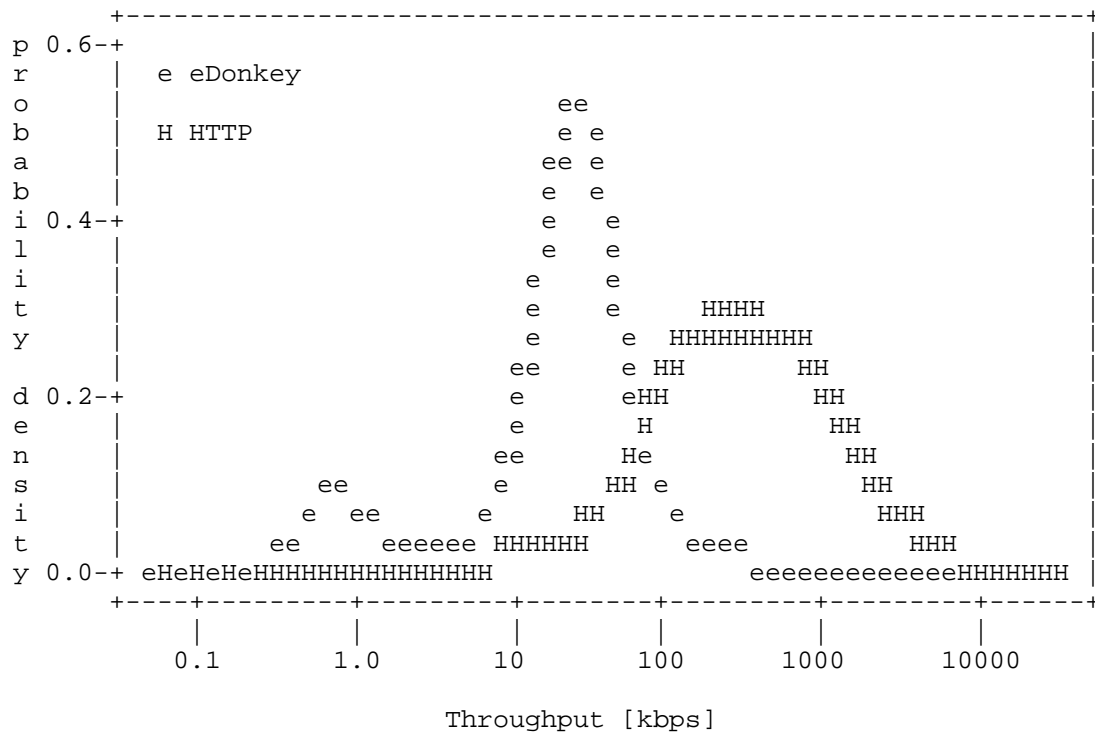


Figure 2 Achieved throughput of flows measured for eDonkey and HTTP applications [37]

Figure 2 displays the throughput distribution of TCP connections for eDonkey peer-to-peer and HTTP applications. It only considers single flow with a length of more than 50 Kbyte. But typically, a web browser uses two to three TCP connections at the same time and an eDonkey client about 10. Still, the throughput of a single HTTP flow is in about an order faster than the of eDonkey flow. In [37], the authors assume this is due to the fact that peer-to-peer connections fill the uplink and that HTTP is used at the faster downlink.

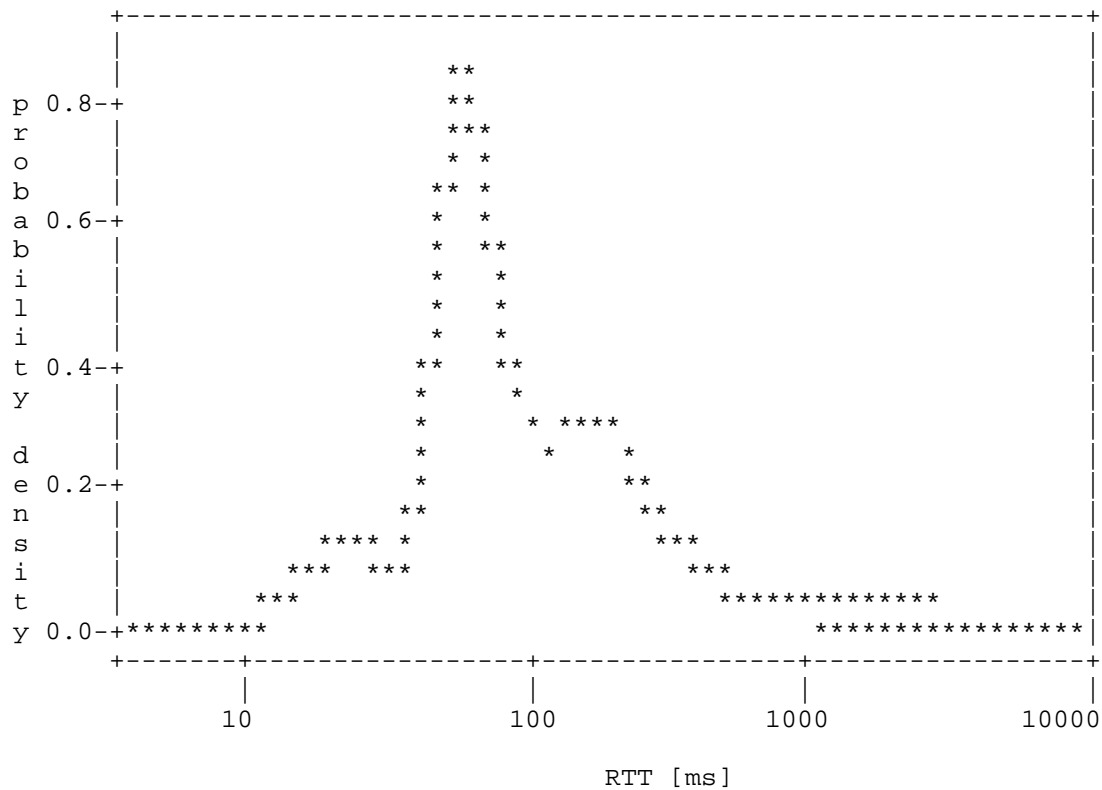


Figure 3 TCP roundtrip times [36]

Figure 3 displays TCP roundtrip times including both access and backbone network. Both graphs can be seen as an indication for the assumption that an application, even in modern Internet access networks, might be subjected to a wide variability of throughput ranging from a few kbits/s up to 10 Gbit/s and TCP round trip times from 5ms up to one of several seconds.

Albeit these results are only valid for TCP, similar results should be expected for RTP over UDP - with a small advantage because UDP flows are not always responsive.

As a summary, a codec for the Internet should be able to work under these widely varying transmission conditions and should be tested against a wide distribution of expected throughputs.

8.4. Transmission Variability on the Internet

Besides effects such as route flapping or link failures modeled in G.1050 [20], the Internet experience in short-time scales sharp changes sharply in bandwidth utilization. For example, [49] and [38] showed that variability of Internet traffic comes in form of spike like traffic increments. Similarly, [32] studied why the Internet is bursty in time scales of between 100 and to 1000 milliseconds.

In the light of these results, one can assume that the IIAC's transmission conditions will vary in similar time scales. More precisely, it will be subjected to

- . variability due to bursty traffic having a duration of between 100 and 1000 milliseconds,
- . interruptions due to temporal link failures every minute to every hour that might have a temporal interruption from 64 ms to several seconds [20], and
- . route flap events every minute to every hour that have a delay of between 2 and 128 ms [20].

8.5. The Effects of Transport Protocols

Realtime multimedia is not always transported over RTP and UDP. Sometimes it makes sense to use a different transport protocol or an additional rate adaptation. The reasons for that are manifold.

- . If a scalable codec shall be supported, RTCP-based feedback information can be utilized to implement a rate control mechanisms [41]. However, RTCP-based feedback suffers from the drawback that RTCP messages are allowed only every 5 s. Thus, implementing a fast responding mechanism is not possible.
- . In the presence of restricted firewalls, VoIP can sometimes only be transmitted over TCP. In those cases, the transmission scheduling is not given by the codec but by TCP. TCP algorithms typically don't have a smooth sending rate but frequently send packets in bursts and change the amount of packets sent every round trip time (Figure 4). More precisely, TCP causes the sending schedule to behave in the following way:
 - . During the Slow Start phase (for example at the beginning of a TCP connection) the transmission rate increases exponentially.

- . If a TCP segment is not acknowledged after about four RTTs, the TCP sending rate starts at one packet per RTT again.
- . During congestion avoidance, the sending rate increases steadily by one segment per RTT.
- . If a congestion event is then detected, the sending rate is reduced by 50%.

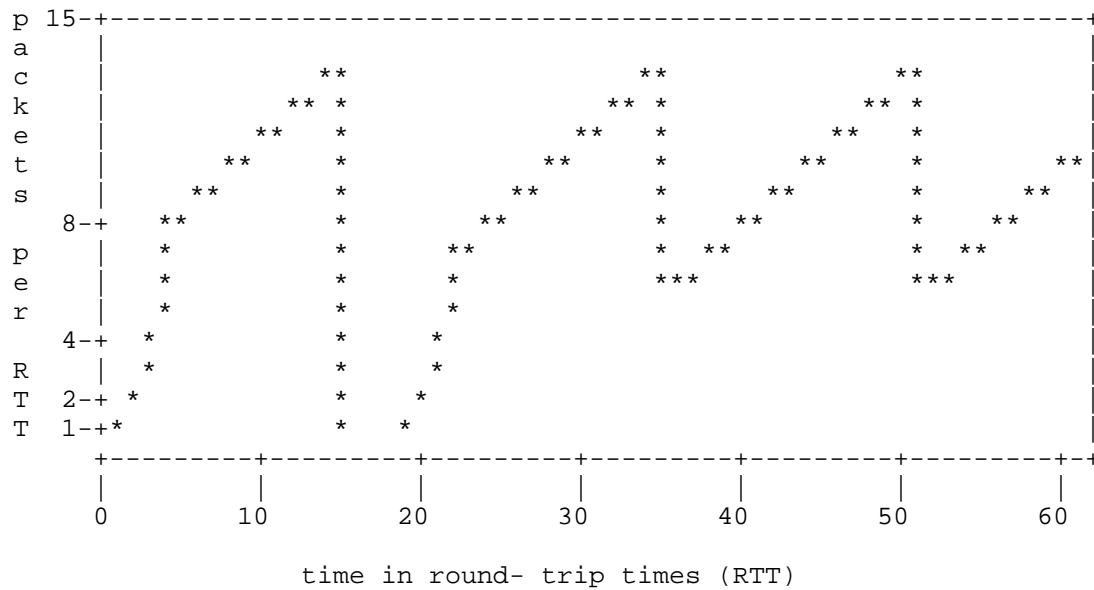


Figure 4 Sending rate of a standard TCP over time

- . The DCCP transport protocol supports multiple congestion control protocols and gives means to support TCP friendliness without retransmission. Thus, it is suitable for real time multimedia transmissions. DCCP supports a TCP emulation, which shows a similar rate over time as TCP, and the TFRC congestion control, which changes its rate in a smoother way (Figure 5). Besides TFRC, which is intended to transmit packets of maximal size (aka MTU), TFRC-SP is optimized for flows with variable packet sizes such as VoIP. With TFRC-SP, smaller packets can be transmitted at a faster pace than it is the case for larger packets because they contribute less to the gross bandwidth consumption. The TFRC protocol might provide a lower bandwidth and a lower QoE as UDP or TCP, unless if not proper optimizations are taken (see [48]). Also, it is suggested to limit the rate control to 100 packets per second. This limit might be too low for an IIAC.

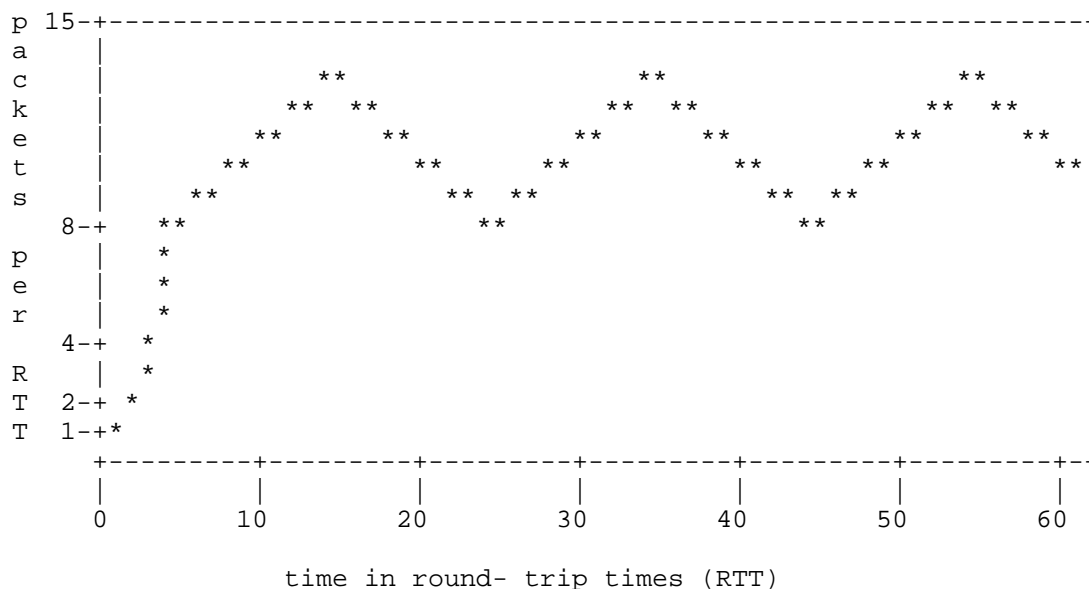


Figure 5 Sending rate of the TFRC protocol

In general, the transport protocol has a clear influence on the transmission conditions. Coding rates need to be adapted by sharply and smoothly to changed bandwidth estimations. Changes of the bandwidth estimation may occur every RTT. Also, in cases of a TCP timeout, the transmission is halted and the decoding must be stalled.

8.6. The Effect of Jitter Buffers and FEC

Both jitter buffers trade frame losses against delay. In cases of a jitter buffer, frames are delayed before playout. This helps in cases of lately arriving frames that would otherwise be ignored and would have to be concealed. Jitter buffers are adaptive and are changing dynamically to the current loss process on the Internet.

Forward Error Correction helps to cope with isolated losses as redundant speech frames are transmitted in the following packets. In the presence of loss, FEC increases the delay because the receiver has to wait for the following packets. Both delay and packet losses are important contributors to the overall Quality of Experience [2].

Since the delay process on the Internet often comes in the form of a gamma distribution, thus a statistical monitor of past delays helps to predict the size of future jitter. Then, if the playout schedule does not match the predicted loss process, playout can be accelerated or slowed down.

However, due to the reasons described in Section 8.4 not all increments in transmission time might be predictable. This has a profound effect on the jitter buffer as it actually cannot predict well, whether a frame is lost or whether it is going to be delayed. If a frame is scheduled for playout but has not been received, the jitter buffer has to consider two cases. First, the frame is lost and has to be concealed. This typically means that the audio signal needs to be extrapolated or interpolated to conceal the gap due to a lost frame. Second, the frame is delayed and shall be played out at a later point in time. Then, the resulting gap in playout must be concealed by extrapolating the previous audio signal.

These issues have an effect on testing the concealment algorithm of the codec. The same concealment function must be tested against time gap concealment and loss concealment.

8.7. Discussion

Judging a codec performance using a realistic model of a transmission channel is difficult. Good models of IP transmission channels are available. However, before a codec can be tested against those channels, further building blocks such as the transport protocol, the jitter buffer, and FEC should be known - at least roughly.

Alternatively, a codec can be tested only against of packet loss patterns only without considering any rate adaption or playout

rescheduling. But then again, the codec should be additionally tested for those impairments, which occur due to the dynamics of the Internet. These include

- o slowing down and speeding up the playout in cases of moderate rescheduling of playout times,
- o stalling and resuming the playout in cases of temporal link outages,
- o moderately reducing and increasing bit and frame rates during contention periods, and
- o sharply reducing (in case of congestion) and fast increasing (during connection establishment) of bit and frame rates.
- o Time gap and loss concealment.
- o Speeding up and slowing down the playout speed.

9. Usage Scenarios

Quality of Experience is the service quality perceived subjectively by end-users (refer to Section 2) and as ITU-T document G.RQAM [21] states "overall acceptability may be influenced by user expectations and context". Thus, in this section we describe the usage scenarios, in which the IIAC codec will probably be used, and the expectations users have in those communication contexts. We list seven main scenarios and describe their quality requirements.

9.1. Point-to-point Calls (VoIP)

The classic scenario is that of the phone usage to which we will refer in this document as Voice over IP (VoIP). Human speech is transmitted interactively between two Internet hosts. Typically, besides speech some background noise is present, too.

The quality of a telephone call is traditionally judged by subjective tests such as those described in [24]. The ACR scale used in MOS-LQS sometimes might not be very suitable for high quality calls, then - for example - the MUSHRA [16] rating can be applied.

A telephone call is considered good if it has a maximal mouth-to-ear delay of 150 ms [17] and a speech quality of MOS-LQS 4 or above. However, interhuman communication is still possible if the mouth-to-ear delay is much larger.

The effect of delay jitter might not be very well notable in case of speech. Thus, playout rescheduling can happen often take place.

In many cases, phone calls are made between mobile devices such as mobile phones and cellular phone. In these cases, energy consumption is crucial and both complexity and transmission rate may be reduced to save resources.

9.2. High Quality Interactive Audio Transmissions (AoIP)

In this scenario we consider a telephone call having a very good audio quality at modest acoustic one-way latencies ranging from 50 and 150 ms [17], so that music can be listened to over the telephone while two persons are talking interactively.

While delay expectations might be similar to those of classic telephony, the audio quality must meet similar standards as those of consumer Hifi equipment like MP3 and CD players, iPods, etc.

If music is played, playout rescheduling events may be heard easily be heard as the rhythm changes. Only a few studies such as [10] have been made to examine the effect of time varying delays on service quality. In general, it can be assumed that the requirements regarding constancies of playout schedules are higher than in case of speech because human beings can notice rhythmic changes easily. Thus, in the presence of music, frequent playout rescheduling shall be avoided.

9.3. High Quality Teleconferencing

Also, for today's teleconferencing and videoconferencing systems there is a strong and increasing demand for audio coding providing the full human auditory bandwidth of 20 Hz to 20 kHz. This rising demand for high quality audio is due to the following reasons:

- o Conferencing systems are increasingly used for more elaborated presentations, often including music and sound effects which occupy a wider audio bandwidth than that of speech. For example, Web conferences such as WebEx, GoToMeeting, Adobe Acrobat Connect are based on an IP based transmission.
- o The new "Telepresence" video conferencing systems, providing the user with High Definition video and audio quality, create the experience of being in the same room by introducing high quality media delivery (such as from Cisco).

- o The emerging Digital Living Rooms are to be interconnected and might require a constant high quality acoustic transmission at high qualities.
- o Spatial audio teleconference solutions increase the quality because they take advantage of the cocktail-party effect. By taking advantage of 3D audio, participants can be identified by their location in a virtual acoustic environment and multiple talkers can be distinguished from each other. However, these systems require stereo audio, if the spatial audio is rendered for headphones.

9.4. Interconnecting to Legacy PSTN and VoIP (Convergence)

This scenario does not include the use case of using a VoIP-PSTN gateway to connect to legacy telephone systems. In those cases, the gateway would make an audio conversion from broadband Internet voice to the frugal 1930's 3.1 kHz audio bandwidth.

The quality requirements in this scenario are low because legacy PSTN typically uses narrow-band voice. Also, in those cases one might expect the codec negotiation might decide on a common codec both for PSTN and VoIP in order to avoid transcoding.

However, the complexity requirements might be stringent because central media gateways must scale to a high number of users. In this context, hardware costs are an important criterion and the codec has to operate efficient.

9.5. Music streaming

Music streaming typically does not require low delays. However, in special cases such as live events and in the presence of alternative transmission technologies, low-delay streaming may be demanded.

Examples are important sport events, which are streamed both on terrestrial, (analogue) and low delay broadcast networks and on IP-based distribution networks. The latter ones becomes aware (such as when a footballer scores) more lately than the ones their neighbors using terrestrial technology.

9.6. Ensemble Performances over a Network

In some usage scenarios, users want to act simultaneously and not just interactively. For example, if persons sing in a chorus, if musicians jam, or if e-sportsmen play computer games in a team together they need to communicate acoustically.

In this scenario, the latency requirements are much harder than for interactive usages. For example, if two musicians are placed more than 10 meters apart, they can hardly stay synchronized. Empirical studies [10] have shown that if ensembles play over networks, the optimal acoustic latency is at around 11.5 ms with a targeted range from 10 to 25 ms.

Also, the users demand very high audio quality, very low delay and very few events of playout rescheduling.

9.7. Push-to-talk like Services (PTT)

In spite of the development of broadband access (xDSL), a lot of users do only have service access via PSTN modems or mobile links. Also, on these links the available bandwidth might be shared among multiple flows and is subjected to congestion. Then, even low coding rates of about 8 kbps are too high.

If transmission capacity hardly exists, one can still degrade the quality of a telephone call to something like a push-to-talk (PTT) like service having very high latencies. Technically, this scenario takes advantage of bandwidth gains due to disruptive transmission (DTX) modes and very large packets containing multiple speech frames causing a very low packetization overhead.

The quality requirements of a push-to-talk like service have hardly been studied. The OMA lists as a requirement of a Push-to-talk over cellular service a transmission delay of 1.6 s and a MOS values of above 3.0 that typically should be kept [39]. However, as long as an understandable transmission of speech is possible, the delay can be even higher. For example, [39] allows a delay of typically up to 4 s for the first talk-burst. Also, [39] describes a maximum duration of speaking. If a participant speaking reaches the time limit, the participant's right to speak shall be automatically revoked.

If the quality of a telephone call is very low, then instead of listening-only speech quality the degree of understandability can be chosen as performance metric. For example, objective tests of the understandability use automatic speech recognition (ASR) systems and measure the amount of correctly detected words.

In any case, the participant shall be informed about the quality of connection, the presence of high delays, the half-duplex style of communication, and its (limited) right to speak. For example this can be achieved by a simulated talker echo.

9.8. Discussion

The requirements of the usage scenarios are summarized in the following table.

Scenario	Sound Quality			Latency			Complexity	
	low	avg.	hifi	10ms	150ms	high	low	high
VoIP	X				X		X	X
AoIP		X	X		X			X
Conference		X			X			X
Convergence	X				X		X	X
Streaming		X	X			X		X
Performances			X	X				X
Push-To-Talk	X					X	X	X

Figure 6 Different requirements for different usage scenarios

10. Recommendations for Testing the IIAC

The IETF IIAC differs substantially from a classic narrow and wideband codec. Thus, the previously applied codec testing procedures such as ITU P.830 cannot be entirely adopted. Instead, one must check carefully, which of the procedures are used without changes, which procedures are used with minor changes and which procedures are dropped or replaced.

In Section 1 we listed five groups of stakeholders, which have different requirements and demands on how to test the quality of an IIAC. In the following, we recommend testing procedures for those stakeholders.

10.1. During Codec Development

The codec development is an innovative process. In general, innovation and research in general benefits from openness and discussion between experts. Thus, format restrictions on how to test the codec might hinder the codec development because innovation may also take place in testing procedures. Instead, many experts both in codec development and codec usage shall be able to participate. If this is the case, they contribute with their expertise, identify weaknesses, and discuss potential codec enhancements. During innovation, openness in participation and discussion is very fruitful and leads to good results.

Based on the ongoing experience, codec developers know best on how to tests their codecs. Typically, those tests include informal

testing, semiformal testing, and expert interviews. They are intended to find weaknesses in the codec, to identify artifacts or distortions, and to achieve algorithmic progress.

10.2. Characterization Phase

The characterization phase is intended to study the features, the quality tradeoff and the properties of a codec under standardization. It is intended to be an objective measure of the codec's quality to convince third parties of the quality properties of the standardized codec. In order to achieve this aim, a formal testing procedure has to be established.

In general, we recommend to base the procedure of the characterization phase on procedures that are similar to those that were used for the G.719 standardization (Section 7.2 and especially [35]). In the following, we describe the suggested testing procedure in the characterization phase.

10.2.1. Methodology

The testing of sound quality can be done using the MUSHRA tests with eight samples and three anchors. One anchor is the known reference, the second one is a hidden reference, and the third one the hidden anchor. It is suggested to use a bandwidth filtered signal with at low-pass filter at 3.5 kHz. However, because a will range of qualities are to be tested ranging from Hifi down to toll quality, it is beneficial to add a further low quality anchor such as a 3.5 kHz bandwidth sample distorted by modulated noise (MNRU) [25], for example with MNRU of a strength of $Q=25$ dB that corresponds to a MOS value of 1.79 [4].

10.2.2. Material

Reference samples should be 48 kHz sampled, stereo channel material. The nominal bandwidth of the reference samples shall be limited to the range of 20 to 20000 Hz. Three different kinds of contents shall be tested: speech, music and mixed content.

Speech samples shall include different languages including English and tonal languages. The speech samples shall be recorded in a quiet environment without background noise or reverberations. The speech samples shall contain one meaningful sentence having a length of about 4 s.

Music samples shall contain a wide variety of music styles including classical music, pop, jazz, and single instruments. The length of

samples shall be of between 10 and 15 s. A smoothing of 100 ms both at the beginning and at the end shall be conducted, if required.

Mixed content may contain advertisements, film trailers, news with jingles and other mixtures of speech, music and noises. The length may be at about 5-6 s.

10.2.3. Listening Laboratory

Multiple independent laboratories shall conduct the listening tests. They are responsible for generating or selecting reference samples as well as for the pre and post screening of subjects. In the end, the results of about 24 experienced listeners shall be published (in addition to the samples).

The tests must be conducted in a quiet listening environment at about NC25 (approximate 35 dBA). For example, an ISOBOOTH room can be used.

It is recommended to use a high quality D/A, such as Benchmark DAC, Metric Halo ULN-2, Apogee MiniDAC. High quality headphone amplifiers and playback level calibration shall be used. Playback levels might be measured via Etymotic in-ear microphones. Also, high quality headphones (e.g. AKG 240DF, Sennheiser HD600) are advisable.

10.2.4. Degradation Factors

The IIAC is likely to be highly configurable. However, due to time limits, only a few parameter sets can be tested subjectively. Thus, we recommend to do subjective studies with

- o different bit rates (from low to high, 5 tests)
- o different frame rates (from low to high, 2 tests)
- o different loss pattern (G.1050 profile A, B, and C at low rate with speech content and at high rate with music content. The influence of jitter, delay, and link failures shall be ignored. In total, this would be 6 tests)
- o different sample contents
 - o Speech, speech+reverberations, and speech+noise+reverberations at low and medium rates (3 tests).
 - o The speech sample must be tested in different languages (English, Chinese, ...) and with male/female voices (6 tests)

- o Mixed content and music shall be tested at medium and high rates (about 10 tests).
- o A low complexity mode, DTX and the FEC mode shall be tested at low rates because they are typically used on constraint devices (3 tests)
- o Abrupt changes in bit and frame rates (reduction by half, exponential start, 2 tests)
- o Smooth changes of bit and frame rates (incrementing or decreasing the codec's gross rate by 1.5 kbyte every 100ms, 2 tests)
- o Stall and continue operations (20, 200, and 1000 ms, 3 tests)
- o Accelerated and slowed down playout (+- 10% for speech at low rates)
- o Reference codecs such as LAME MP3, G.719, and AMR each at two coding rate (6 tests)

Already, these are 48 different tests that need to be conducted.

In addition, for intermediate values objective tests shall be run using PEAQ (for music) and P.OLQA (for speech). The intermediate results shall be mapped on the MUSHRA scale with a quadratic regression because PEAQ and P.OLQA are using an ODG and MOS scale respectively.

10.3. Application Developers

Application developers can take advantage of the results of the qualification phase. They may use the results to develop a quality model, which describes the expected quality of the codec at a given parameter set (refer to [11] for an example).

In addition, they can test their system using the draft G.1050 simulation model, which is especially useful for optimizing rate control, dejittering buffers and concealment algorithms. Different systems may be tested with quality models, subjective listening tests, conversational listening tests, or with objective measures such as POLQA.

Also, field tests may be conducted to test the effect of a real network on the VoIP application.

10.4. Codec Implementers

To tests the conformance of a codec, codec implementers can use objective tools like PEAQ or P.OLQA to see, whether the newly implemented codec performs in a way that is similar to the performance of the reference implementation. These tests shall be done for many different parameter sets.

10.5. End Users

End user may be included in the qualification tests. The intentions of these tests are two-fold. First, the awareness of the end-user shall be increased. Second, querying users may be a cost effective way of conducting listening-only tests.

However, before the rating results of end users can be considered for further usage, one need to compare between formal and web-based testing results to see, to what extent they differ from each other.

11. Security Considerations

The results of the quality tests shall be convincing. Thus, special care has to be taken to make the tests precise, accurate, repeatable and trustworthy.

Some testing houses may have a conflict of interest between accurate quality ratings and promotion of own codecs. Thus, a high degree of openness shall be enforced that requires all of the testing material and results to be published. This way, others may verify the results of testing houses. In addition, some stimuli shall be tested by all the testing houses to compare their quality of rating.

Moreover, hidden anchors may help to identify subjects, which rate the quality of samples less precisely.

12. IANA Considerations

This document has no actions for IANA.

13. References

13.1. Normative References

13.2. Informative References

- [1] R. Birke, M. Mellia, M. Petracca, D. Rossi, "Understanding VoIP from Backbone Measurements", IEEE INFOCOM 2007, 26th IEEE International Conference on Computer Communications, pp.2027-2035, May 2007.
- [2] C. Boutremans, J.-Y. Le Boudec, "Adaptive joint playout buffer and FEC adjustment for Internet telephony," IEEE Societies INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications., vol.1, pp. 652- 662 vol.1, 30 March-3 April 2003.
- [3] Broadcom, "BCM1103: GIGABIT IP PHONE CHIP", Jan. 2005, <http://www.datasheetcatalog.org/datasheet2/3/07ozspx224dsarq6zu13i2ofyqyy.pdf>
- [4] N. Cote, V. Koehl, V. Gautier-Turbin, A. Raake, S. Moeller, "Reference Units for the Comparison of Speech Quality Test Results", Audio Engineering Society Convention 126, May 2009.
- [5] Ericsson, "Analysis of PEAQ's applicability in predicting the quality difference between alternative implementations of the G.722.1FB coding algorithm", ITU-T SG12, Received on 2008-05-09, Related to question(s) : Q9/12, Meeting 2008-05-22.
- [6] ETSI TC-TM, "ETR 250: Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3,1 kHz handset telephony across networks", ETSI Technical Report, July 1996.
- [7] S. Floyd, E. Kohler, "Profile for Datagram Congestion Control Protocol (DCCP) Congestion ID 4: TCP-Friendly Rate Control for Small Packets (TFRC-SP)", RFC 5622, August 2009.
- [8] S. Floyd, E. Kohler, "TCP Friendly Rate Control (TFRC): The Small-Packet (SP) Variant", RFC 4828, April 2007.
- [9] J. Gruber, G. Williams, Transmission Performance of Evolving Telecommunications Networks, Artech House, 1992.

- [10] M. Gurevich, C. Chafe, G. Leslie, S. Tyan, "Simulation of Networked Ensemble Performance with Varying Time Delays: Characterization of Ensemble Accuracy", Proceedings of the 2004 International Computer Music Conference, Miami, USA, 2004.
- [11] C. Hoene, H. Karl, A. Wolisz, "A perceptual quality model intended adaptive VoIP applications", International Journal of Communication Systems, Wiley, August 2005.
- [12] J. Holub, J.G. Beerends, R. Smid, "A dependence between average call duration and voice transmission quality: measurement and applications," Wireless Telecommunications Symposium, 2004, pp. 75- 81, May 2004.
- [13] ITU, "Incoming LS: Proposed G.1050/TIA-921B IP Network Model Simulation", ITU-T SG 12, Temporary Document 268-GEN, May 12, 2010.
- [14] ITU, "ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", Recommendation, October 1997.
- [15] ITU, "ITU-R BS.1387: Method for objective measurements of perceived audio quality", Recommendation, November 2001.
- [16] ITU, "ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems", Recommendation, January 2003.
- [17] ITU, "ITU-T G.107: The E-model: a computational model for use in transmission planning", Recommendation, April 2009.
- [18] ITU, "ITU-T G.114: One-way transmission time", Recommendation, May 2003.
- [19] ITU, "ITU-T G.191: Software tools for speech and audio coding standardization", Recommendation, March 2010.
- [20] ITU, "ITU-T G.1050: Network model for evaluating multimedia transmission performance over Internet Protocol", Recommendation, November 2007.
- [21] ITU, "ITU-T G.RQAM, "Reference guide to QoE assessment methodologies", standard draft TD 310rev1, May 2010.

- [22] ITU, "ITU-T P.10/G.100: Vocabulary and effects of transmission parameters on customer opinion of transmission quality", Recommendation, July 2006.
- [23] ITU, "ITU-T P.800: Methods for objective and subjective assessment of quality", Recommendation, August 1996.
- [24] ITU, "ITU-T P.805: Subjective evaluation of conversational quality", Recommendation, April 2007.
- [25] ITU, "ITU-T P.810: Modulated noise reference unit (MNRU)", Recommendation, February 1996.
- [26] ITU, "ITU-T P.830: Subjective performance assessment of telephone-band and wideband digital codecs", Recommendation, February 1996.
- [27] ITU, "ITU-T P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", Recommendation, February 2001.
- [28] ITU, "ITU-T P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO", Recommendation, November 2003.
- [29] ITU, "ITU-T P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs", Recommendation, November 2007.
- [30] ITU, "ITU-T P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2", Recommendation, November 2007.
- [31] ITU, "ITU-T P.880: Continuous evaluation of time-varying speech quality", Recommendation, May 2004.
- [32] H. Jiang, C. Dovrolis, "Why is the internet Traffic Bursty in Short Time Scales?" Sigmetrics'05, Banff, Alberta, Canada, June 2005.
- [33] C. Lamblin, R. Even, "Processing Test Plan for the ITU-T G.722.1 fullband extension optimization/characterization phase", ITU-T Study Group 16, Temporary Document TD 322 (WP 3/16), 22 April - 2 May 2008.

- [34] C. Lamblin, R. Even, "G.722.1 fullband extension characterization phase test results: objective (ITU-R BS.1387-1) and subjective (ITU-R BS.1116) scores", ITU-T Study Group 16, Temporary Document TD 341 R1 (WP 3/16), 22 April - 2 May 2008.
- [35] C. Lamblin, R. Even, "G.722.1 fullband extension optimization/characterization Quality Assessment Test Plan", ITU-T Study Group 16, Temporary Document TD 323 (WP 3/16), 22 April - 2 May 2008.
- [36] J. Lee, J. Kim, C. Jang, S. Kim, B. Egger, K. Kim, S Han, "FaCSim: A Fast and Cycle-Accurate Architecture Simulator for Embedded Systems", in Proceedings of the International Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES'08), Tucson, Arizona, USA, June 2007, Software available at <http://facsim.snu.ac.kr/>.
- [37] G. Maier, A. Feldmann, V. Paxson, M. Allman, "On Dominant Characteristics of Residential Broadband Internet Traffic", IMC'09, November 4-6, 2009, Chicago, Illinois, USA.
- [38] T. Mori, S. Naito, R. Kawahara, S. Goto, "On the characteristics of internet traffic variability: Spikes and Elephants", SAINT'04, 2004.
- [39] Open Mobile Alliance, "Push to talk over Cellular Requirements", Approved Version 1.0, 09 Jun 2006, OMA-RD-PoC-V1_0-20060609-A.pdf
- [40] OPTICOM, SwissQual, TNO, "Announcement of OPTICOM, SwissQual and TNO to submit a joint P.O.LQA model", ITU-T SG 12, Contribution 117, Received on 2010-05-07. Related to question(s): Q9/12.
- [41] D. Sisalem, A. Wolisz, "Towards TCP-friendly adaptive multimedia applications based on RTP", IEEE International Symposium on Computers and Communications, pp. 166-172, 1999.
- [42] S. Smirnoff, K. Pupkov, "SoundExpert, How it Works, Audio quality measurements in the digital age", <http://soundexpert.org/>, revived Nov. 2010.
- [43] L. Sun, "Speech Quality prediction For Voice Over Internet", PhD thesis, University of Plymouth, January 2004, <http://www.tech.plymouth.ac.uk/spmc/people/lfsun/mos/>.

- [44] Texas Instruments, "C64x+ CPU Cycle Accurate Simulator", October 2010,
http://processors.wiki.ti.com/index.php/C64x%2B_CPU_Cycle_Accurate_Simulator.
- [45] Texas Instruments, "TNETV3020: Carrier Infrastructure Platform, Telogy Software products integrated with TI's DSP-based high-density communications processor", 2008,
<http://focus.ti.com/lit/ml/spat174a/spat174a.pdf>
- [46] TransNexus, "Asterisk V1.4.11 Performance", webpage, accessed Nov. 2010,
http://www.transnexus.com/White%20Papers/asterisk_V1-4-11_performance.htm
- [47] K. Vos, K. Vandborg Sorensen, S. Skak Jensen, J. Spittka, "SILK", presentation at the 77th IETF meeting in the WG Codec, March 22, 2010, Anaheim, USA.
<http://tools.ietf.org/agenda/77/slides/codec-3.pdf>
- [48] H. Vlad Balan, L. Eggert, S. Niccolini, M. Brunner, "An Experimental Evaluation of Voice Quality Over the Datagram Congestion Control Protocol," IEEE INFOCOM 2007. 26th IEEE International Conference on Computer Communications. pp. 2009-2017, 6-12 May 2007.
- [49] J. Wallerich, A. Feldmann, "Capturing the Variability of Internet Flows Across Time", Proceedings INFOCOM 2006. 25th IEEE International Conference on Computer Communications, 23-29 April 2006.
- [50] M. Westerlund, "How to Write an RTP Payload Format", work in progress, draft-ietf-avt-rtp-howto-06, Internet-draft, March 2, 2009.
- [51] Wikipedia contributors, "Bit rate", Wikipedia, The Free Encyclopedia, 10 October 2010, 20:00 UTC,
http://en.wikipedia.org/w/index.php?title=Bit_rate&oldid=389931944
- [52] Wikipedia contributors, "Cycle accurate simulator", Wikipedia, The Free Encyclopedia, 4 September 2010, 14:27 UTC,
http://en.wikipedia.org/w/index.php?title=Cycle_accurate_simulator&oldid=382876676

- [53] Wikipedia contributors, "Latency (engineering)", The Free Encyclopedia, 15 October 2010, 23:54 UTC,
[http://en.wikipedia.org/w/index.php?title=Latency_\(engineering\)&oldid=390971153](http://en.wikipedia.org/w/index.php?title=Latency_(engineering)&oldid=390971153)
- [54] Wikipedia contributors, "Profiling (computer programming)", Wikipedia, The Free Encyclopedia, 15 August 2010, 03:57 UTC,
[http://en.wikipedia.org/w/index.php?title=Profiling_\(computer_programming\)&oldid=378987422](http://en.wikipedia.org/w/index.php?title=Profiling_(computer_programming)&oldid=378987422).
- [55] M. T. Yourst, "PTLsim: A cycle accurate full system x86-64 microarchitectural simulator", in ISPASS '07, 2007, software available at <http://www.ptlsim.org/>.

14. Acknowledgments

This document is based on many discussions with experts in the field of codec design, quality of experience and quality management. My special thanks go to Michael Knappe, Sebastian Moeller, Raymond Chen, Jack Douglass, Paul Coverdale, Jean-Marc Valin, Koen Vos, Bilke Ullrich, and all active participants of the Codec WG mailing list. Also, I like to express my appreciation to the members of the ITU-T study groups 12 and 16, with whom I had many fruitful discussions.

Authors' Addresses

Christian Hoene
Universitaet Tuebingen
WSI-ICS
Sand 13
72076 Tuebingen
Germany

Phone: +49 7071 2970532
Email: hoene@uni-tuebingen.de

