Subjective audio tests: quality of a selection of codecs when used in IP network

Catherine COLOMES (*), Martin VARELA (**), Jean-Charles GICQUEL (*)
(*) France Telecom R&D, DIH/EQS/MAI, 4 rue du Clos Courtel, 35512 Cesson-Sévigné, France
Phone: +33 2 99 12 44 53, Fax: +33 2 99 12 37 13
{catherine.colomes, jeancharles.gicquel}@francetelecom.com
(**) IRISA – INRIA Rennes, Campus universitaire de Beaulieu, 35042 Rennes, France
Phone: +33 2 99 84 71 37, Fax: +33 2 84 25 29
martin.varela@irisa.fr

Subjective listening tests were performed in order to evaluate the impact of transmission over an IP network on the quality of two "well-known" audio codecs. Those two codecs were WM9 and Helix 9 (Real), tested at two different modes and bit-rates.

A dedicated network simulator was used in collaboration with the IRISA institute (Rennes 1 university), who developed this tool. This network simulator is accurately described in the paper. The network perturbations considered were packet losses (the parameters used were the loss rate and the mean burst size), delay and jitter. All these four parameters were adjusted independently allowing for different combinations of network parameter values. Many configurations were generated and a large amount of time was spent in order to listen carefully to the different impact on the overall quality of the codecs. Eventually, six configurations that yielded six distinctly different quality levels were kept to run the quality tests. The kept parameters values are commonly found in actual IP networks.

As the goal of the audio tests was to assess the quality, the subjective audio test methodology used was the one known as MUSHRA. That stands for MUlti Stimuli with Hidden Reference and Anchor points. This is a method dedicated to the assessment of intermediate quality. It has been recommended at the ITU-R under the name BS.1534 [1]. An important feature of this method is the inclusion of the hidden reference and specific signals as anchor points along the quality scale. The objective of the test was to observe the "quality behaviour" of the two mentioned codecs at two different bit rates in a simulated network environment. Consequently, the anchor points were the two codecs at a given bit rate without any simulated network, and the full band reference.

Results presented in the paper are of interest not only because they show the specific weakness of the two codecs when faced with network perturbations but also because they show a specific use of the MUSHRA method, which makes these kinds of tests possible.

Key words: Network simulation, subjective audio tests, MUSHRA

1 Introduction

The objective of these tests was to evaluate the impact of an IP network on the quality of some audio codecs. Tests were performed by simulating the network in order to have complete control of the sound test sequences and the network parameters.

The goal was to judge the quality of the Windows Media Players V9 (WM9) and Helix Real One V9 (H9) codecs at different bit rates and modes simulating the impact of an Internet network-type transmission.

This paper reports the results of these tests. It contains detailed description of the network simulation performed, the tested codec configurations, the audio excerpts used, the testing process, the statistical analysis and the results.

2 Audio codecs under test

The following 2 audio codecs were tested:

- Helix / Real 9 (commercial solution)
- Windows Media 9 (commercial solution)

Here is a brief description.

2.1 Helix / Real 9: Helix Producer Plus 9

Helix producer Plus Version 9 Audio codec is built from ATRAC3 Sony technology bought in 2000 by RealNetworks. Audio codecs are the same as those in version 8 and 8.5. However now, surround audio codecs are available.

2.2 Windows Media 9:

Version 9 of Windows Media was released in September 2002 as a beta version. We expect an evolution in quality of the audio codec compared to that of version 8. As for Helix9, the main change is the availability of a surround codec.

2.3 Encoding parameters

The followings tables give encoding parameters used in those tests for each codec.

Helix 9	**20 kbps[1] mono**	**64 kbps stereo**
Fe / KHz	22.05	44.1
Encoder name	Helix Cook 3	Helix Cook 24 (Ra8)

WMA	**20 kbps mono**	**64 kbps stereo**
Win 9 Fe / KHz	22.05	44.1

Tables 1: Encoding parameters.

3 Test Items

The items were chosen in order to be able to run (later on) audiovisual tests and to compare their results to those of audio tests. That means that the audio items were extracted from sound tracks of audiovisual material.

The items had to be as close as possible to reality (which can be found in a network service), bearing in mind that they also had to remain as critical as possible.

The 5 used audio excerpts are listed in table 2:

Item name	**Description**
Basket	Speech commentaries (male voice) of a basket ball match with applause and people singing and shouting
James Bond	Male and female speech and car race with a lot of stereo effects.
Jazz	Jazz music with a female singer
Canoe	Speech commentaries (male voice) with some classical music in the background
Film	French movie trailer "Aller simple". Speech, music, different kinds of noises.

Table 2: test items

Duration of the items was about 15 seconds but in order to avoid buffering side effects, the items were looped in order to get a 1 minute sequence per item.

[1] Kbps : kilo bits per second

4 Network simulation

The coded files were generated with the latest version of WM9 and Helix9, on a PC. This PC is the server in the simulation of an Internet connection. The network simulator was located in between the PC server and the workstation from where the requests were sent as shown on figure 1.

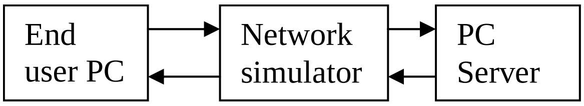

Figure 1: Network simulation

This work was done in collaboration with the IRISA2 at Rennes 1 University. The "ARMOR" team headed by Mr Rubino lent us the network simulator that will be described hereafter.

4.1 The network simulator

Mr Varela who works in the "ARMOR" team built this network simulator based on different references:

- "Characterizing End-to-End Packet Delay and Loss in the Internet", J-C. Bolot, Journal of High-Speed Networks, December 1993, vol.2, n°3.
- "Analysis and Control of Audio Packet Loss over Packet-Switched Networks", Bolot, J-C. and Crépin, H., Technical report, INRIA, 1993.
- "ACM Multimedia Systems", The case for {FEC-based} error control for packet audio in the Internet, Bolot, J-C. and Vega Garcia, A., 1996.
- "Capacity of a Burst--loss Channel", Gilbert, E., Bell Systems Technical Journal, September 1960, vol.5, n°39.

The idea behind this network simulator prototype is to affect flows traversing it in the same way as they would be affected by a big network such as Internet. In order to achieve this, packet losses are introduced, and the packets are delayed by a (normally slightly) variable amount of time.

The implementation is based on the Linux Kernel fire walling subsystem. Rules are defined to send the relevant packets to a user space application (the simulator) for processing. The simulator allows the definition of channels that identify flows. These channels are defined by rules much like those used for packet filters, and can be unidirectional or bi-directional. For each channel, there are 4 parameters to configure, namely the loss rate, the mean loss-burst size, the mean delay, and a value for the jitter in the form of a percentage of the delay.

Arriving packets are matched against the defined channels, and if they don't match, they are let through. If they do match, the simulator decides if the packet is to be dropped or let through. The packet loss model used is based on the "Gilbert' model" (see reference above) although it is simpler than this one. It consists on a two-state homogeneous Markov chain, where in state 0 packets are accepted, and in state 1, packets are dropped. Transition from state 0 to state 1 happens with a probability p and implies a loss: transition from state 1 to state 0 happens with a probability q. It is straightforward that:

$$p = \frac{1}{MBS_m} \frac{PLR_m}{1 - PLR_m}$$

and $$q = \frac{1}{MSB_m}$$

Where PLR is the packet loss rate and MBS is the mean size of the loss bursts. This model has been extensively used in literature, and it has been shown to mimic closely the loss processes found on the Internet.

If the packet is let through, it is queued for an amount of time depending on the mean delay and jitter specified for the channel. We are not aware of any suitable jitter model for IP networks in the literature, so this implementation uses exponentially distributed variation, centred on the mean delay, which makes for a pretty strong jitter. The formula used to calculate the actual delay is:

$$D=d-j+Exp(j)$$

Where d is the mean delay, j is the jitter and D is the amount of time by which the packet is actually delayed. Processed packets are placed in a queue ordered by exit time which is regularly checked and if any packet is ready to be re-injected in the network, the packet is de-queued and a message is sent to the kernel so that the packet is accepted by the firewall. As a side note, only the packet headers are copied to user space, as the payload is of no interest for our purposes.

In order to provide a good real-time performance, the simulator should run on a dedicated machine with no other services active.

[2] Institut de Recherche en Informatique et Systèmes Aléatoires

4.2 The network configurations

Twenty-six different network configurations were generated. The parameters values for those configurations are listed in table 3.

N° Configuration	Packet loss rate	Mean loss bursts size	Mean delay	Jitter
C1	6 %	1.2	150 ms	15 ms
C2	15 %	2.7	30 ms	15 ms
C3	15 %	1.2	30 ms	30 ms
C4	10 %	2	150 ms	5 ms
C5	6 %	2.7	200 ms	15 ms
C6	10 %	1.2	250 ms	5 ms
C7	10 %	2.7	100 ms	15 ms
C8	6 %	2.7	150 ms	5 ms
C9	10 %	1.2	200 ms	30 ms
C10	15 %	1.2	100 ms	5 ms
C11	15 %	2	150 ms	30 ms
C12	10 %	2.7	200 ms	30 ms
C13	6 %	2	150 ms	15 ms
C14	6 %	1.2	200 ms	15 ms
C15	2 %	1.2	30 ms	15 ms
C16	10 %	2.7	200 ms	15 ms
C17	10 %	2.7	150 ms	15 ms
C18	2 %	2.7	100 ms	5 ms
C19	15 %	2	200 ms	30 ms
C20	6 %	1.2	100 ms	30 ms
C21	2 %	2	30 ms	15 ms
C22	2 %	2	100 ms	15 ms
C23	6 %	1.2	200 ms	5 ms
C24	2 %	2.7	200 ms	5 ms
C25	15 %	2	200 ms	15 ms
C26	10 %	2	200 ms	30 ms

Table 3: Network configurations

Those 26 configurations are representative of different kinds of links that can be found on a big network such as the Internet. They are based on real measurements, in order to be able to reproduce what really happens on IP networks. Those measurements are usually performed by non-intrusive probes.

Six main sets of different qualities were found from those 26 configurations. That led to a selection of only 6 network configurations that are representative of all configurations. Only those six configurations were kept to test the two mentioned codecs in order to lighten the subjective tests. The choice was made after a pre-listening of the generated defaults of all configurations. This pre-scanning was made by expert listeners. The selected ones are listed in table 4.

N° Configuration	Packet loss rate	Mean loss bursts size	Mean delay	Jitter
C1	6 %	1.2	150 ms	15 ms
C2	15 %	2.7	30 ms	15 ms
C3	15 %	1.2	30 ms	30 ms
C5	6 %	2.7	200 ms	15 ms
C6	10 %	1.2	250 ms	5 ms
C15	2 %	1.2	30 ms	15 ms

Table 4: Selected Network configurations

Then, for each of those six network configurations, the helix9 and WM9 compressed items were played through the simulated network.

4.3 Settings of codecs

With regard to the codecs parameters in relation with the network, it was decided to fix some of them in order to avoid settings in favour of one of the codecs. That led to the following settings:

- The buffer length was fixed to 20 seconds;
- The connexion configuration was that of a DSL wire with a bit-rate of 384 kbps.
- The packets loss was applied on all packets (TCP, UDP,) in an identical way for both codecs, and on the up link as well as the down link.
- On the helix player the "turbo play" option was off. This option allows a faster bufferisation and it doesn't exist on the Windows player.

4.4 The files recording

In order to generate the wave files, when not directly possible with the players, we used the Total Recorder 4.0 software. This software redirects the players output to a virtual soundcard that is a wave file. Then, it records uncompressed audio stream on the hard disk in wave format. We made sure that the silences (usually due to strong packet loss) were not removed from the final recording.

Next, 15 seconds were taken out from this 1 minute recording per audio excerpts and tested configurations in order to run the tests. We decided to choose the worst case (pre listening that had been made by experts) between second 15 and the end of the 1 minute file.

The following step consisted of the synchronisation of the coded audio files to the reference wave files. This was done using automatic software to calculate the delay in number of samples and Cool Edit Pro to synchronise the files.

5 Test Process

5.1 Test method

The MUSHRA methodology was used for the quality test. MUSHRA stands for MUlti Stimuli with Hidden Reference and Anchor points. This is a method dedicated to the assessment of intermediate quality. It has been recommended at the ITU-R under the name BS.1534 [1]. This was developed in 1999 by the EBU Project Group B/AIM in collaboration with the ITU-R Working Party 6Q. An important feature of this method is the inclusion of the hidden reference and bandwidth limited anchor signals.

For this special test, we chose to work with different anchors points. The objective of the test was to observe the "quality behaviour" of the 2 mentioned codecs at 2 different bit rates in a simulated network environment. Consequently the anchor points were the two codecs at a given bit rate without any simulated network, and the full band reference.

5.2 Training phase

Each listener had a training period of about 15 mn, in order to become familiar with the test methodology and software and with the kind of quality they had to assess. This was also an opportunity to adjust the restitution level that would then remain constant during the test phase.

5.3 User Interface

The MUSHRA method has the advantage of displaying all stimuli for one test item at a given bit-rate at the same time. The subjects were therefore able to directly carry out any comparisons between them.

Implementation of MUSHRA user interface from CRC (SEAQ) was used in the test. A screenshot of one implementation of the user interface is shown in figure 2. The buttons represent all the configurations/codecs under test including the hidden reference and both the anchor signals, and the reference, which is specially displayed on the left as "REF". Above each button, with the exception of the "REF" one, a slider is used to grade the quality of the test item according to the continuous quality scale.

For each of the test items, the signals under test were randomly assigned, with a different assignment for each subject. In addition, the test items were randomised for each subject within the session to avoid sequential effects.

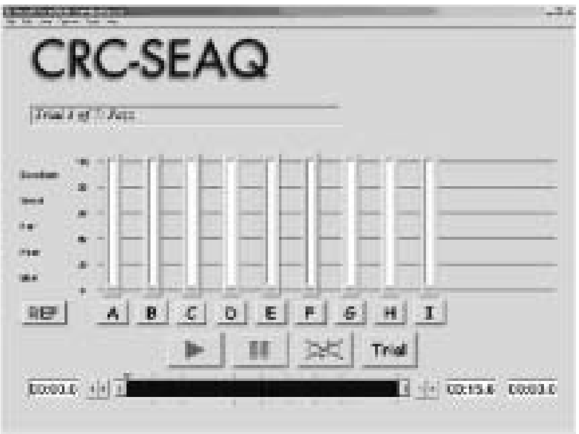

Figure 2: MUSHRA Software

The whole test was divided in two sessions, each containing 3 different network configurations. The first one contains network configurations 1, 3 and 6. And the second one contains network configurations 2, 5 and 15 (refer to table 4 for details).

5.4 The Listening Panel

The listening panel consisted of 18 subjects, most of them experienced in audio but not professionally involved. As some of them didn't follow the test instructions or were not reliable enough (see "post-screening of subjects" in the Statistical Analysis), their scores were discarded. This resulted in a total of 15 reliable listeners.

5.5 Tests instructions and duration

The test instructions explained to the listeners how the software works, what they would listen to (briefly), how to use the quality scale and how to score the different excerpts. It was also an opportunity to mention the fact that there was a hidden reference signal to score, and consequently, there should be at least one score equal to 100 per audio excerpt. This would later be used in the rejection process of listeners.

As there were 6 different network configurations to test, the overall test was split in two sessions. Whatever the session, its duration was between 1 hour and 1 and ½ hour. Every 20 mn, the listener was asked to rest for a while. The training phase was included in this time schedule.

5.6 Listening conditions

The tests were performed on the STAX Signature SR-404 (open)3 headphones and their SRM-006t amplifier. The subjects had the possibility to set the reproduction level individually before they started the actual test (during the training phase). The subjects were then restricted from changing the reproduction level during the test.

The test items were stored on a Windows 2k workstation. The digital sound was played through the PC board Digigram VX 222 and converted by a 24 bits DAC (3Dlab DAC 2000).

ITU-R has defined specific requirements for the listening conditions to ensure comparable and reliable results of subjective assessments of sound systems [2]. This covers:

- the acoustical characteristics of the listening room and the sound field therein,
- the arrangement of the monitoring loudspeakers in the listening room,
- the location of the listening positions for the test.

The listening room used here fulfilled the majority of the corresponding requirements, taking into account that tests were performed on headphones. Other important tests had previously been performed in this room, for example MPEG tests, 3GPP tests, EBU tests, European projects tests, etc..

6 Statistical Analysis

6.1 General analysis

The statistical analysis method described in the MUSHRA specifications was used to process the test data. The results are presented as mean grades and 95% confidence intervals.

Experience has shown that the scores obtained for different test sequences are dependent on the criticality of the test material used. Therefore, these figures have been included in this report in order to provide a more complete understanding of codec performance. This is done by presenting results for different test sequences separately rather than as aggregated averages across all the test sequences used in the assessment.

[3] http://www.son-video.com/Rayons/Hifi/Casques/Stax.html

6.2 Post-screening of subjects

Two post-screening methods were proposed:

- One was based on the ability of a subject to make consistent repeated grading;
- The other relied on inconsistencies of an individual grading compared with the mean result of all subjects for a given item. This was done by looking at the individual spread and the deviation from the mean grading of all subjects. The aim of this was to obtain a fair assessment of the quality of the test items.

When "intermediate" quality is tested, a subject should be able to easily identify the reference signal and the coded version. In addition, a subject should be able to give a grade that corresponds to the grade given by the majority of the subjects. Subjects with grades at the upper end of the scale are likely to be less critical. Subjects who have grades at the lowest end of the scale are likely to be too critical. The methods are primarily used to eliminate subjects who cannot make the appropriate discriminations.

The easiest way to measure the inconsistencies of an individual subject compared to the mean result is to calculate the correlation coefficient. This coefficient $\rho_{x,y}$ is used to determine the relationship between 2 sets of data. It is calculated as follows:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x . \sigma_y}$$

Where:

$-1 \leq \rho_{x,y} \leq 1$ And

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

In which *n* is the total number of listeners, *X* is the set of scores given by listener *x*, and *Y* is the set of average scores given by the *n* listeners and:

$\sigma_x = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}(x_i - \mu_x)^2}$ (Respectively with the *Y* set)

$\mu_x = \frac{1}{n}.\sum_{i=1}^{n} x_i$ (Respectively with the *Y* set)

Consequently, subjects whose coefficient $\rho_{x,y}$ was below 0.88 were discarded. The fact that the hidden reference was found is taken into account in the rejection process, resulting in the rejection of 3 listeners:

1. One listener was not critical enough: 8 scores were equal or above 90, different from the hidden references; in addition, his grading was inconsistent compared with the mean result: $\rho_{x,y}$ <0.88;
2. One listener has scored one hidden reference 75 and 21 scores were equal or above 90, different from the hidden references (among them, 4 scores equal to 100); $\rho_{x,y}$ <0.88;
3. One listener has scored two hidden references 90 and 92 and 8 scores were equal or above 90, different from the hidden references (among them, 2 scores equal to 100); $\rho_{x,y}$ <0.8;

7 Results

7.1 Mono Mode 20 kbps

7.1.1 Global results

Figure 3 shows the global results obtained for the network configuration C1, C3 and C6 (referring to table 4) on all the 5 audio excerpts and for both codecs in mono mode.

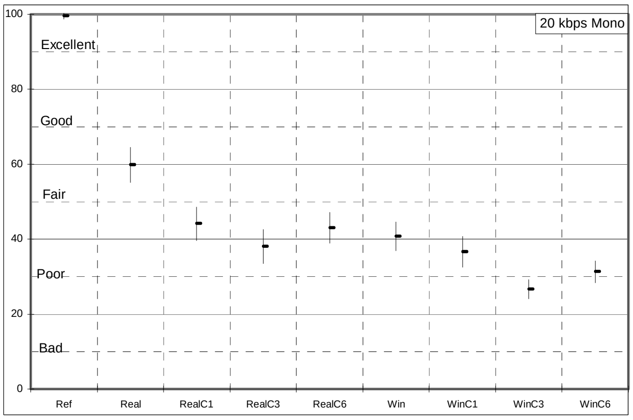

Figure 3: Global results on C1, C3 and C6 at 20 kbps mono.

Figure 4 shows the global results obtained for the network configuration C2, C5 and C15 (referring to table 4) on all the 5 audio excerpts and for both codecs in mono mode.

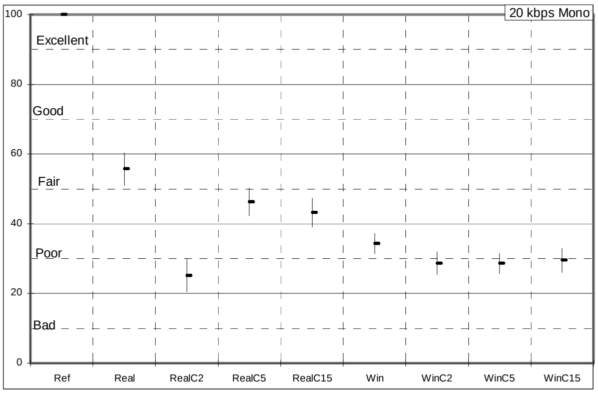

Figure 4: Global results on C2, C5 and C15 at 20 kbps mono.

The different tested configurations are spread along the X-axis while the quality scale is along the Y-axis (labels go from "excellent" = [100-80] to "bad" = [20-0]).

As expected, the hidden reference items are rated the highest with a very small confidence interval if any.

It is noticeable that the scores given to the anchor codecs (with no network simulations) are lower in the second test when configurations C2, C5 and C15 have been tested. The difference is around 5 points on the quality scale and there is a slight overlapping between the confidence intervals. This difference can be explained by a difference in quality between those 2 tests: the overall quality of the test with C2, C5 and C15 was slightly better than that of the C1, C3 and C6 test. This tends to score the anchor codecs lower.

Now, looking at configuration by configuration, we can observe a difference in quality between the two codecs Windows Media 9 and Helix Real 9.

1. For configuration C1 (figure 3), the Real codec scored 45, at the bottom scale of the "Fair" quality range while the Windows codec scored 35, at the upper level of the "Poor" quality range. In this configuration, the packet loss rate is somewhat low (6%) with a mean loss burst size of 1.2, while the mean delay and the jitter are important (respectively 150 ms and 15 ms). The result is that the Win codec is slightly more sensitive to the delay and the jitter than the Real one.

2. For configuration C2 (figure 4), the Real codec scored 25, at the lower level of the "Poor" quality range while the Windows codec scored 29, at the middle of the "Poor" quality range. In this configuration, the packet loss rate is important (15%) with a high mean loss burst size of 2.7, while the mean delay is small (30 ms), and the jitter is quite high at 50% (15 ms). The result is that both codecs are sensitive to the packet loss rate and mean loss burst size.

3. For configuration C3 (figure 3), the Real codec scored 38, at the upper level of the "Poor" quality range while the Windows codec scored 27, at the lower level of the "Poor" quality range. In this configuration, the packet loss rate is important (15%) with a mean loss burst size of 1.2, while the mean delay is less important (30 ms) and the jitter is higher than that of the 2 previous configurations (30 ms). The result is that the Win codec is slightly more sensitive to the jitter than the Real one.

4. For configuration C5 (figure 4), the Real codec scored 47, at the lower end of the "Fair" quality range while the Windows codec scored 28, at the lower level of the "Poor" quality range. In this configuration, the packet loss rate is rather low (6%) with a high mean loss burst size of 2.7, while the mean delay is very important (200 ms) and the jitter is equal to 30 ms. Once again, the result is that the Win codec is more sensitive to delay than the Real one.

5. For configuration C6 (figure 3), the Real codec scored 44, at the lower end of the "Fair" quality range while the Windows codec scored 32, at the middle of the "Poor" quality range. In this configuration, the packet loss rate is important (10%) with a mean loss burst size of 1.2, while the mean delay is the most important of all configurations (250 ms) and the jitter is very low (5 ms). Once again, the result is that the Win codec is more sensitive to delay and packet loss size than the Real one.

6. For configuration C15 (figure 4), the Real codec scored 44, at the lower end of the "Fair" quality range while the Windows codec scored 30, at the middle of the "Poor" quality range. In this configuration, all parameters have low values compared to the previous 5 configurations: the packet loss rate is quite low (2%) with a mean loss burst size of 1.2, while the mean delay is equal to 30 ms and the jitter is 15 ms. Once again, the result is that the Win codec is more sensitive to network parameters than the Real one.

To conclude, we can say that on the average, the quality of Win codec at 20 kbps mono in a simulated network is lower than that of the Real codec for the same configuration, albeit the overall quality is generally speaking rather poor.

An important fact to mention is that the C2 configuration seems to be the most disturbing one for both codecs. This configuration generated a lot of disturbing mutes with the Real codec while there were less mutes with the Win codec. In this case, the Win codec seems to have a better "recall" or "rebuilding" strategy than the Real one.

7.1.2 Student T test

The next tables (tables 5 and 6) show the results of a Student T test between both codec at the same network configuration.

STUDENT	Real	RealC1	RealC3	RealC6
Win	0,00			
WinC1		0,02		
WinC3			0,00	
WinC6				0,00

Table 5: Student T Test results at 20 kbps mono for configurations C1, C3 and C6.

STUDENT	RealC2	RealC5	RealC15
WinC2	**0,22**		
WinC5		0,00	
WinC15			0,00

Table 6: Student T Test results at 20 kbps mono for configurations C2, C5 and C15.

Figures calculated by a Student T test are the probability that two compared configurations are significantly different or not in quality (intersection between a line and a column).

In our case, this test was used to observe whether the quality of the Real codec for a specific network configuration was significantly different from that of the Win codec for the same network configuration.

The following assumptions were made in order to calculate tables 5 and 6:

- The Student T test uses the bilateral distribution;
- The T test was done over two sets of samples with different standard deviation;

In the obtained results, a number higher than 0.05 means that the two compared codecs are not statistically different in quality.

Looking to the results of table 6, the bold figures show that for the network configuration C2, the Real and Win codecs are not statistically different from a quality point of view. Following the remarks made in the previous section, that means that this network configuration has a very strong impact on the quality of both codecs. They cannot cope with a high packet loss rate (15%) and a high mean loss bursts size (2.7) – referring to table 4.

7.1.3 Results per audio excerpts

Figures 5 and 6 show the results obtained for each tested configuration for all 5 audio excerpts. The average over the 15 listeners and the confidence interval at 95% are displayed.

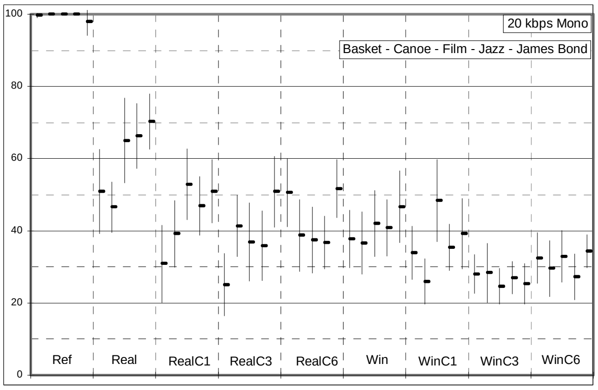

Figure 5: Results for all the audio excerpts on C1, C3 and C6 at 20 kbps mono.

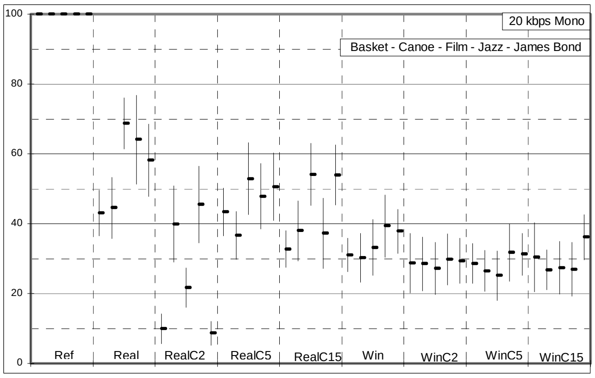

Figure 6: Results for all the audio excerpts on C2, C5 and C15 at 20 kbps mono.

For one given configuration (column), all the 5 audio excerpts are ranked in the same order, that is:

1. Basket
2. Canoe
3. Film
4. Jazz
5. James Bond

Usually, these kinds of results are used to see if some audio items are more critical than other. Here, the only thing that can be said is that the influence of mutes (for the real codec in configuration C2) is very high on the "Basket" and "James Bond" excerpts.

7.2 Stereo Mode 64 kbps

7.2.1 Global results

Figure 7 shows the global results obtained for the network configuration C1, C3 and C6 (referring to table 4) on all the 5 audio excerpts and for both codecs in stereo mode.

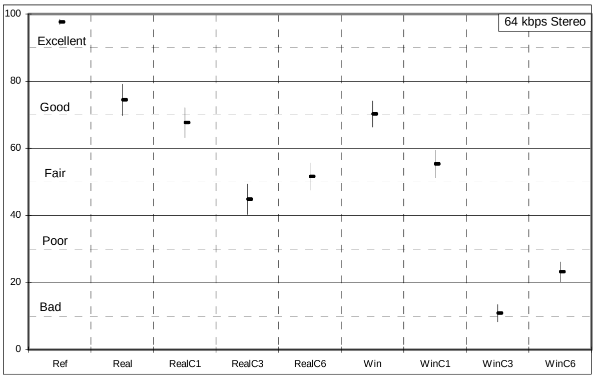

Figure 7: Global results on C1, C3 and C6 at 64 kbps stereo.

Figure 8 shows the global results obtained for the network configuration C2, C5 and C15 (referring to table 4) on all the 5 audio excerpts and for both codecs in stereo mode.

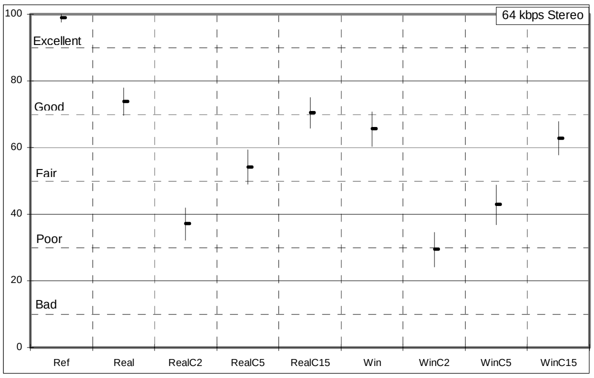

Figure 8: Global results on C2, C5 and C15 at 64 kbps stereo.

The different tested configurations are spread along the X-axis while the quality scale is along the Y-axis (labels go from "excellent" = [100-80] to "bad" = [20-0]).

As expected, the hidden reference items are rated the highest with a very small confidence interval.

It is noticeable that the score given to the anchor codec Real (with no network simulations) is identical for both tests, while the one given to the anchor codec Win is lower in the second test when configurations C2, C5 and C15 have been tested. The difference is around 5 points on the quality scale and there is a slight overlapping between the confidence intervals. This is not a major difference.

The global quality of those anchor codecs is higher than that of the mono tests which is reassuring.

Now, looking at configuration by configuration, we can observe a difference in quality between the two codecs Windows Media 9 and Helix Real 9.

1. For configuration C1 (figure 7), the Real codec scored 68, at the middle of the "Good" quality range while the Windows codec scored 56, at the upper level of the "Fair" quality range. In this configuration, the packet loss rate is quite low (6%) with a mean loss burst size of 1.2, while the mean delay and the jitter are important (respectively 150 ms and 15 ms). The result is that the Win codec is slightly more sensitive to the delay and the jitter than the Real one.
2. For configuration C2 (figure 8), the Real codec scored 37, at the upper level of the "Poor" quality range while the Windows codec scored 30, at the middle of the "Poor" quality range. In this configuration, the packet loss rate is important (15%) with a high mean loss burst size of 2.7, while the mean delay is less important (30ms) and the jitter is relatively high (15 ms). The result is that both codecs are sensitive to the packet loss rate and mean loss burst size.
3. For configuration C3 (figure 7), the Real codec scored 45, at the lower end of the "Fair" quality range while the Windows codec scored 11, at the middle of the "Bad" quality range. In this configuration, the packet loss rate is important (15%) with a mean loss burst size of 1.2, while the mean delay is less important (30 ms) and the jitter higher than that of the 2 previous configurations (30 ms). The result is that the Win codec is more sensitive to higher jitter values than the Real one.
4. For configuration C5 (figure 8), the Real codec scored 55, at the upper level of the "Fair" quality range while the Windows codec scored 43, at the lower level of the "Fair" quality range. In this configuration, the packet loss rate is somehow low (6%) with a high mean loss

burst size of 2.7, while the mean delay is very important (200 ms) and the jitter is equal to 30 ms. Once again, the result is that the Win codec is more sensitive to the delay than the Real one.

5. For configuration C6 (figure 7), the Real codec scored 52, at the middle of the "Fair" quality range while the Windows codec scored 24, at the lower end of the "Poor" quality range. In this configuration, the packet loss rate is important (10%) with a mean loss burst size of 1.2, while the mean delay is the most important of all configurations (250 ms) and the jitter is very low (5 ms). Once again, the result is that the Win codec is more sensitive to the delay and the packet loss size than the Real one.
6. For configuration C15 (figure 8), the Real codec scored 70, at the middle of the "Good" quality range while the Windows codec scored 63, at the lower end of the "Good" quality range. In this configuration, all parameters have low values compared to the previous 5 configurations: the packet loss rate is quite low (2%) with a mean loss burst size of 1.2, while the mean delay is equal to 30 ms and the jitter is 15 ms. This configuration has the lowest impact on both codecs compared to that of the other configurations.

To conclude, we can say that on the average, the quality of Win codec at 64 kbps stereo in the simulated network is lower than that of the Real codec for the same configuration. Nevertheless, the overall quality at 64 kbps stereo is higher than that at 20 kbps mono which is expected.

However, there is still this remark about the influence of the C2 configuration that once again seems to be the most disturbing one for both codecs. This configuration generated a lot of disturbing mutes and both codecs in stereo mode were affected.

7.2.2 Results per audio excerpts

The next tables (tables 7 and 8) show the results of a Student T test between both codec at the same network configuration.

STUDENT	Real	RealC1	RealC3	RealC6
Win	0,04			
WinC1		0,00		
WinC3			0,00	
WinC6				0,00

Table 7: Student T Test results at 64 kbps stereo for configurations C1, C3 and C6.

STUDENT	RealC2	RealC5	RealC15
WinC2	0,04		
WinC5		0,01	
WinC15			0,03

Table 8: Student T Test results at 64 kbps stereo for configurations C2, C5 and C15.

As mentioned in section 7.1.2, numbers calculated by a Student T test are the probability that two compared configurations are significantly different or not in quality (intersection between a line and a column). The same assumptions as in 7.1.2 are made.

As previously, a result higher than 0.05 means that the two compared codecs are not statistically different in quality.

Looking at the results of table 7, both codecs are statistically different in quality whatever the network configuration.

7.2.3 Results per audio excerpts

Figures 9 and 10 show the results obtained for each configuration tested for all 5 audio excerpts. The average over the 15 listeners and the confidence interval at 95% are displayed.

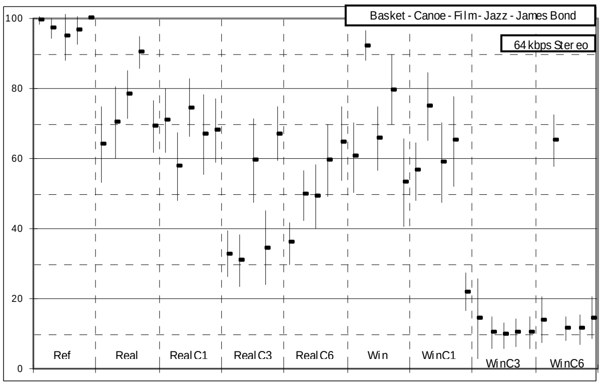

Figure 9: Results for all the audio excerpts on C1, C3 and C6 at 64 kbps stereo.

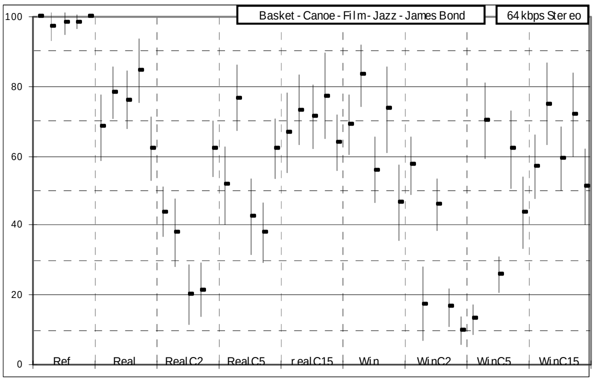

Figure 10: Results for all the audio excerpts on C2, C5 and C15 at 64 kbps stereo.

For one given configuration (column), all the 5 audio excerpts are ranked in the same order, that is:

1. Basket
2. Canoe
3. Film
4. Jazz
5. James Bond

Usually, those kinds of results are used to see if some audio items are more critical than other. It is clear that 3 audio excerpts ("Canoe", "Film", "Jazz") may be less critical than the others as their average score is lower than 100 and they have a non-zero confidence interval.

8. Conclusion

Considering the results, it is obvious that the network configuration C2 makes the overall quality of both codec Real and Microsoft decrease whatever the mode and bit rate. This is quite a "strong" configuration with a high packet loss rate (15%) and a high mean loss bursts size (2.7) with 30 ms of mean delay and 15 ms of jitter.

Looking at the influence of the other configuration, the main point is that the Win codec seems more sensitive than the Real one, especially at the C3 and C6 configuration, whatever the mode and bit rate. The common feature between those two configurations is a high packet loss rate (15% and 10%). Furthermore, the mean delay of C6 configuration is very high (250 ms). Finally, it seems that the WM9 codec is more sensitive to high packet loss rate and mean delay than the Real codec.

[1]: ITU-R Recommendation BS.1534 “Method for the subjective assessment of intermediate quality level of coding systems” Geneva (June 2001).

[2]: ITU-R Recommendation BS.1116 “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems” Geneva (1994).