Minutes of the VQEG MM and JRG-MMQA meeting in Seoul, 19-22 October 2004

Participants in person:
Pierre Brétillon (PB), TDF
Alex Bourret (AB), BT
Kjell Brunnström (KB), Acreo
Dave Hands (DH), BT
Quan Huynh-Thu (QH), PsyTechnics LTD
Kyung-Mee Kim, RRL
Chulhee Lee (CL), Yonsei
Jun Okamoto (JO), NTT
Margaret Pinson (MP), NTIA/ITS
Eugen Rodel (ER), SwissQual Inc
Christian Schmidmer (CS), Opticom
Osama Sugimoto (OS), KDDI
Akira Takahashi (AT), NTT
Arthur Webster (AW), NTIA/ITS
Jae-seung Kim Samsung
DongHwan Kim Samsung

On telephone:
Viviak Balasubrawmanian (VB), Intel Corp
Phil Corriveau (PC), Intel Corp
Irina Cotanis (IC), Ericsson

On Internet:
2 anonymous persons

Tuesday 19 October 2004

- All the participants introduced themselves
- Agenda approved for the whole week. Some people are leaving about lunchtime on Friday so the goal is to get done by then.
- AW sent the agenda and current testplan out to the reflector.
- DH gives an update of the status of the work since the Rome meeting.

Pre-test reports:

- NTIA: AW presented the results from his pre-test. He will provide a short summary for the minutes before the end of the week. Subjects could tell MPEG4 from reference, but not reference with different frame rates (e.g. generally reference ar 25 fps rated equivalent to reference at 12.5 fps). QH comments that the subjects are not trying to tell the different sequences apart, that is trying to see the difference, rather than scoring the quality. There might be just slight difference in quality, even if they can see that they differ. MP comments that it was good that the test was short. Some subjects were annoyed that the whole screen was not filled. Something to think about for the instructions for our tests. The question comes up whether the target application should be in the instructions or not. PC says that the subjects should be instructed about the application. Viewing distance was not controlled, but was about 8H.
- BT: AB and DH are presenting their pre-test results. DH will provide a summary of the results for the minutes. MPEG 4 had better scores than H.263. The mouse position affected the viewing distance. The

variance was quite high. Std dev 0.5. Most of the scores for the encoded sequences are in the lower quality range (typically MOS= 2 or less). MP points out that we need to be careful in setting up the test to get the middle qualities in the test. QH agrees with that. BT used a modified version of mplayer, which they can make available to everyone. They use Labview to control the experiment. AW comments that it is dangerous to have a still or nearly still sequence, since subjects might mark them low just because they are still. CS offers to make available an MPEG4 filter. AW and DH will be sharing their subjective scores with others. This will be sent to the MM pretest reflector.

- NTT presented their input documents. They present a product developed at NTT ATC, which costs about $13000 for capturing and displaying video Okamoto (2004a). It captures the digital image that is about to go out to the display. The pre-test results are described in Okamoto (2004b).

- SwissQual is presenting their preparation of test (Rodel, 2004a) material and the results from the pre-test (Rodel, 2004b). MP points out it might be to specialized for VQEG if we go for this method. There is an issue however going from VBR to CBR, which may introduce distortions due to interpolation.

- Yonsei gives an overview of their pretest. They have compared for a subset of the data the DSCQS and ACR (both with and without HRR). They have rated the 200 clips. The results will be presented later. CL fears that models might do very differently on different data sets.

- QH presented his input document for the meeting, see Huynh-Thu (2004). The model basing the first comment in this document, based on SwissQual has not been discussed by the group. MP points out that it might not be the same upsampling and not upsampling in the general case. CL says that is not. QH ask what is part of the test condition. MP says that it should be the same for the subject and the objective model. AW think that it is not any point to use point 2 or 3 in the SwissQual model. QH does not agree and think that point 3 is the most important

  - VB email:

  I understand the issue is what should be done when a HRC is taking in a full frame rate (25/30fps) reference video and is outputting a processed video that is of much lower frame rate (8/12.5/15fps).

  1. Though the HRC is doing the downscaling from 25 to 12.5 (for instance), for comparison purposes, you ideally need to have both the original and processed video at the same frame rate so that the comparison is done apples to apples. My first suggestion was, if you can isolate the preprocessing step (the downscaling of frame rate) from the HRC, you can apply the same pre-processing to the reference sequences to convert them to the same frame rate as the processed for comparison. So in effect, you will be comparing a pre-processed reference with the processed video. If the preprocessing step cannot be isolated, we can use a consensus "reference/golden" scaler implementation that will do the pre-processing of the originals. I don't believe that adding another processing step (up scaling between point 3 and 4) is a good idea since, you don't want any additional variables in your test path. How would you know if the degradation you see in the processed video is being caused by the HRC or by the post-processing step? The metric user would be responsible for characterizing the post processing algorithm which is not ideal.
  2. Further, I think, the model should not be compensating for any changes in the frame rate between original and processed videos. A frame rate down sampling that is acceptable for this particular case may not be acceptable for other applications. In my testing, I would like to see any drop in frames be reflected in the scores. My current thinking is that we should have a model score that provides a subjective opinion of all the degradation between the original and processed videos without any automatic compensation of selective artifacts (like frame rate downscaling). References Huynh-Thu, Q. (2004), “Comments and proposals on the actual issues in the MM test plan”, Input document to VQEG meeting Seoul 2004, Psytechnics Limited

- QH continues with the rest of his document. There will not be a discussion now, but rather at the appropriate time during the discussion of the testplan. MP will respond when we discuss the point on subjective test campaign, which will be discussed under the scope of the test. PC points out that we need to instruct our subjects very carefully so we know what their references are. DH is not sure that this is the best way.

- Discussion about the scope of the test. MP think that the size estimated by QH is way too high by a factor of 10.There might be 4 independent labs and 8 proponent labs, which means that the burden could be shared among many labs. There must be a way to check the inter lab correlation. MP has done a test where they done one test and then done another were they have mixed the data from other test and the

correlation was very high. BT could do 4x20 minute test = 400 clips with at least 24 subject. Yonsei could run 1000 clips, SwissQual about 400, PsyTechnics 400, NTT 400, KDDI ?, TDF ?, NTIA 200, Acreo ??.

References

Huynh-Thu, Q (2004), “Comments and proposals on the actual issues in the MM test plan”, Input document for VQEG Seoul meeting 2004, Psytechnics Limited, UK

Okamoto, J. (2004a), “Proposed video capturing system for subjective assessment of VQEG”, Input document for VQEG Seoul meeting 2004, NTT, Japan

Okamoto, J. (2004b), “Results of preliminary tests at NTT and proposed subjective assessment Method”, Input document for VQEG Seoul meeting 2004, NTT, Japan

Rodel, E. (2004a), “MM Pre-test – Impact of different video players (Preparation of the Test Material)”, Input document for VQEG Seoul meeting 2004, SwissQual, Switzerland

Rodel, E. (2004b), “Results of MM Pre-test – Impact of different video players”, Input document for VQEG Seoul meeting 2004, SwissQual, Switzerland

Wednesday 20 October 2004

DH: We did not discuss second email by Vivik. Will be discussed at the appropriate point in the meeting.
Right now we need to specify the type of test we want to do. Psytechnics' document describe 5 type of error that could be tested. We need to work out which can be used in the first test phase.
MP: should take the ILG opinion on this point : the test might end up being too long
DH: if we can specify what we want to do right now (transmission / compression / post processing / combination...) Once this is defined, all the work required can be discussed separately.
MP: could specify the type of HRCs we want to see, and in what proportions. That would be simpler approach.
KB: might define the type of distortions we want to see, without spelling out the combination of codec and condition needed to obtain it.
DH: aim of the test is to get representative range of errors.
DH: Decision can be made whether we want to combine some conditions.
CS: would like to see minimal combinations of errors conditions.
DH : could have compression errors (frame rate / bit rate / codec), transmission errors, decoding errors... Do we want specific test for each ?
AW: First need to test model. If can obtain extra information, good, but not a priority. Would run into problems if one labs do compression and another doing transmission errors, without method to compare -> uge pile of test condition distributed across the labs.
DH: want to identify main categories we want to see in the test.
MP: Live condition was seen as important by some people during the Rome meeting. Should keep them. (Nokia). would be very valuable if they can produce these types of HRCs
CS: Makes decision easier if tests are specific (ex: this model is good at transmission errors, not compression...)
MP: would like to see a "robust test"
KB:
CS: Live condition strong selling point
KB: There are repeatable procedure for test conditions
DH: Are the error going to be very different ?
CS: will know the answer after the test.
CS : were not talking about video conferencing, but live transmission of video.
AW: will have to submit our models prior to the testing. If it takes a year to do the testing, will have to submit the models then. By the end of the test, models can be obsolete.

CS: << procedure for running test and retraining models in P???>>
CL:
CS: Live condition mean using real network and not a simulation
AT: snapshot of real network
Will come back to this to see if we agree to the proposed categories.

Irina's contribution:
3 parts in the document. Paragraph 2 are comments about the current test plan. 2nd part summaries Ericsson's proposal, last paragraph lists the proposal.

KB: Data analysis has not been addressed yet in the testplan. These currently present in the testplan are left over of the previous version.
DH: interesting request for models to output standard MOS value.
AW: in one section, requires minimum perf value for at main measures. This has been controversial in the past. What would be a minimum required value for, say, Pearson correlation value.
Irina : being 85% for speech quality metrics, it could be 90% for video.
KB: Comment from Greg on this contribution. Greg agrees on Irina's comments. Psytechnics also has inputs on the subject of data analysis.
QHT : desirable to simplify scheme for analysing perf. Between polynomial and <<>> : 3rd polynomial could be guaranteed to be monotonic, tools could be made available to vqeg.
Irina : not difficult to have a constrain on these function, but once this is done, you do not allow the function to map efficiently -> arm the algo output.
KB asks Irina if ready to draft a new section on data analysis with Greg to have as an input for the next meeting. Probably will not be able to finalise by the end of the Seoul meeting.
AW: could also review the section produced by the end of this meeting.
Irina agrees.

15 min break

Vivik's second email:
would like to see :
encode / decode errors over a variety of bitrates
simulated internet scenario
would like to see MPEG2 and H264 based HRCs (mpeg2 around 2Mb/s)
HRCs with pre and post processing steps such as 3:2 pull down, deblocking filter, noise reduction filters, PIP scaling.

AW: would be possible to have MPEG2 for 601 test, but not the CIF and QCIF
DH: only 601 has 2Mb/s.
KB: maybe then we also want a CRT display for 601.
Vivik on why PIP is important : we see scaling as preprocessing in the test, wanted to see some has post processing as well.
DH: in current testplan, no scaling allowed.

Return to the pretest:
Suggestion from the floor on conclusion that can be done from the pretest:
MP: value perceived to instruct subjects on what is the best quality they will see during the test, especially in CIF and QCIF tests.
AW: not sure we know which player to use. Some (wmp, powerpoint) have some drawbacks that needs to be overcome. Need to make choices on avi and color space as well.
KB: player could have an impact for high quality have variable frame rate, so player could have impact.
CS: difference was in the preparation of the files.
CL: Handling of transmission errors will be different according to the decoder, so need various decoders
Decide one player or validate a few ?
QHT&AW: should use the same player

MP: chosen player should handle the 601 in term of bandwidth
AB: final player is meant to be transparent, so why specify one ?
CL: 601 can lead to very expensive hardware, so changing might be a problem for labs
DH: we do not seem to have enough information to make a decision right now.
DH: standardise the hardware ?
CL: capture the DVI to capture the output ?
CS: we have to use a refresh rate on the monitor at least twice has much as in the video stream to be able to capture it. Might be a problem for variable display.
KB: using loged frame rate would have impact on many stage of the test, from capture to display.
MP: do Phil or Vivik have rec for any alternative player hanling variable framerate video ?
PC: no rec.
KB: for 601, can use clipstation ?
AB: prob with 525 and 625 resolution, and aspect ratios
CL: software was written to display 601 on LCD.
AB : will try mplayer system with 601
PC: Player as no effect ? so can move on.

AW: not happy with the idea of 1hour test per user, but will go with the group
CS: usually, concentration time about 20 min
DH: will keep the test plan as it is

<<Lunch time>>

Test plan editing (1.4):
MP: include 3-4 Mb/s ? so need to change sentence in the intro

[2.1, ACR.]
DH: If people have contribution with use of ACR and hidden ref removal, would be worthwhile
MP: Until checked with rec 501, would like to keep "with hidden ref removal" in the title
DH: too much time on this. Is "ACR with HRR" a method or not.
DH: in rec501, it is specified
MP: would like to see "ACR-HRR" retained for the rest of the document.
Change accepted.

[general description]
[2.1.2]
MP: might want to reconsider the use of simulated mobile platform. Might be too complicated to setup.
KB: can be interpreted in different way. Using grey around a CIF or QCIF can be seen as simulated mobile platform (not using window around the video).
AW: should define what that means. Should either define the phrase or remove it.
QHT: The rest of the doc will define the type of display, so can be removed from this section of the document.
DH: Can also redefine the scope of the mm test
QHT: put the type of applications
PC: why are we mentioning applications ?
AW: telling the user of the test plan what we're doing
MP: use of ACR for all 3 stages, propose to remove that sentence
AW: big change, would require 2/3
PC : want to remove the sentence too
AW: needs some new knowledge to take any decision, so should be taken later
KB: not sure there was a decision taken about acr method for all 3 stages
QHT: was never discussed.
stage 2 is listed as strictly audio stage, ACR not approved for that.
AW: procedure is : if a test plan is on the reflector, then people can send there comment if they do not agree with it. If not, what's in the test plan is considered as adopted.
No one can remember when the sentence was put in.

3 options :
1 - keep the sentence (0 vote)
2 - remove (6 votes)
3 - replace all with "it is expected that..." (4 votes)
But objection with the vote procedure

vote two :
0 to keep the sentence, 10 to change it
then, 6 to 4 to remove the sentence
-> Sentence removed.

KB: there are several other problems in this paragraph. Viewing distance for example, this would mean for QCIF that the subject is restricted by chinrest (?) or similar.
CP: we have not agreed that chin rest are not the way to go for tests.
DH: chin rest are common in psychological studies

no objection to change "require" to "request"
it would leave the issue open in cases where the subject do not comply in using the chin rest

[display and setup]
"The LCD display should..." should be reworded (grammar).
CP: will there be a list of process given to specific people at the end of the meeting. Up to now : chin rest & LCD setup
The paragraph was intended when 2 labs were planed.
CS: "when combining test results, the brand and model of monitors should be the same"
AW: a little bit too strong, for future statistical use of the data.
KS: the panel is the key thing in a flat panel display, not the brand or ref.
DH: response time of the LCD is important, as shown in the NTT test
DH: Last meeting : opposition to impose given monitor
AW: for broadcast test, specify the characteristics, not the brand. In ITU, "should" means "must". ITU being the target, we need to be careful.
DH: NTT test specifies EIZO CG21 50ms, DELL ultrasharp 16ms
Vivik: sony panel,
DH: do we want a range of monitor to cover the range of panel existing or specifying as much as possible? If we use different panels, need to have a complete desc of their specs. Thinks that if 2 labs are doing a experiment with comparing results in mind, it is important to use the same panel.

The text is replace with "ex:TBD" in the example of panel that can be used.

DH: is there a variability in the specs given?
Note on response time included in the text, but still need to decide the value.
DH: another issue is the post processing used in the LCD, which is unknown.
MP: maybe "it is preferred that everyone uses the same monitor" is preferable
CL: might be interesting to write what is the procedure to know the panel used in the LCD.
CS: write the reason behind the change
MP: If unpractical to get the same monitor, it opens the possibility to still compare the results
QHT: normalisation should compensate for differences between labs
MP: test has shown little impact on score between broadcast monitor and consumer grade LCD
DH: would like to see
Vivik: url for professional lcd monitor
DH: Sony monitor might be interesting to investing, but no need to decide now.

page 8
call for objections to this paragraph "The LCD display..." modifications: no objection, text approved.

AW: do we want to modify "this set up procedure may be..." to allow modification by VQEG, not

VQEG subj. test set-up group?
no objection to remove the all sentence -> sentence removed.

<<20 min coffee break>>

[2.1.4 Viewers]
MP: was suggested to use one order per subject (Psytechnics' contribution). But cannot be forced since might be forced to use tapes-> "it is preferred that each subject...". Number of possible subgroups left opened.
QHT: now too confusing, because proposal is to do full randomisation, regardless of the number of sessions
PC: is it not going to be an organizational nightmare?
DH: problem if using pen and paper, but ok if everything is kept automatically
MP: needs to put a sentence telling that software needs to be provided?
Call for objection on the paragraph change: no objection -> change accepted

next paragraphe, change on the word television-> video
Editor note that definition of non expert viewer is needed
DH: no ITU standard on def. of "expert"?
No objection to the editorial change of that paragraph

AW: normal is "20/30", not 20/20. Also, up to 20/40 there is no statistical difference in the MOS score (GC).
CP wants the data
Will be reviewed at the next meeting

AW: need to define the audio test involved in screening.
section heading changed "Subjects" with 1 objection.

QHT: would like to see some screening for subjects that did not complete the test

[2.1.5 Viewing Conditions]
MP: do we want to say that rec.500 will be followed?
DH: can't remember why we did decide not to.
MP: can't we put the wall coloring ?
DH: Adding that test environment should be quiet and conform to Rec 500
Where possible, viewing condition will comply with P910.
Now Rec.500 changed with P910.
AW: could be stricter and require P.910

No objection to the changes to that paragraph.
No objection to change "where possible" to "laboratories must"

<<end>>

References:
Irina's contribution

Processes / tasks:
chin rest & LCD setup

Thursday 21 October

Discussion of Contributions

Presentation of KDDI's contribution (VQEG Introduction of MM test sequences.pdf) by OS:

KDDI produced multiview test sequences (e.g. 9 cameras with 320x240 each) and provides them on FTP site (ftp.kdilabs. ????). It is proposed to use only one view point for MM. Duration of sequences is between 20 and 30s.

QH: Upsampling is bad idea.

DH: Open source?

OS: for VQEG free, but not open source.

QH: Sequences he had seen were already compressed. Leads to discussion of source format and how to provide it. OS said that the uncompressed ones can be made available. No compression was applied. The images were shot at 320x240 pixels. No details on cameras.

Note by DH:NTT said that they can test up to 1000 clips. This increases the possible test size significantly

Presentation on pretest by CL:

Comparison of ACR MOS and ACR DMOS. All data showed some outliers. The analsis is still ongoing.

ACR MOS	ACR DMOS	R=0.9507
ACR MOS	DSCQS MOS	R=0.910
ACR DMOS	DSCQS DMOS	R=0.9126

Discussion started on analysis and interpretation of pretest, but due to ongoing analysis, it was hard to come to further conclusions. CL would like to see other labs performing further experiments.

Further work on test plan:

2.1.6. Test Data Collection

Discussion on who is responsible for data collection (ILG, proponents or both). Sentence will be changed to reflect responsibility of co-chairs and proponents. Also the procedure of collection is specified now.

AW: Idea is also that only those who contribute can get access to the data.

MP: format of data files must be agreed.

→ MP will create example table and to describe format. Probably still during this meeting. PC will contribute.

→ New text was explicitly agreed upon with no objections.

New sections:

2.2. Data Format

2.2.1 Results Data Format

Format of the results of subjective tests. Tbd.

2.2.2 Subject Information Data Format

Format of the data with information on the subjects used (viewer number etc…)

Long discussion on weather to keep data of subjects which were rejected.

MP: proposal, collecting is fine, but do not submit to VQEG

DH: proposes text: Subjective test laboratories will collect the following information from subjects: viewer number, age, gender, visual acuity, colour vision test results

It was noted, there may be some country specific rules regarding privacy, like no permission to record names etc.

Discussion will be continued. Text is not yet approved.

<<Coffee break>>

2.2.3. Subjective Data Analysis

Entire section was widely edited to reflect change from DSCQS to ACR with hidden reference removal. Additional changes as follows:

Exclusion of subjects that missed a vote is more relevant for pen-and-paper tests than for fully automated tests.

Discussion on whether or not multiple subjects at a time may perform the test. There may be problems with randomisations.

AW: Can be overcome by forming subgroups.

DH: Intention was not to ensure that only subject per display is allowed. Proposal to adjust text accordingly.

CL: Concern is if test of subjects is not synchronised, they may disturb each other (e.g. one is watching while another one is writing).

New text : "…only one subject *per display* assessing …"

Change accepted with no objections

Section 2.3.1 of Rec 500 is the one to be used to screen subjects. "Screening for DSIS DSCQS and alternative " To be done by each lab.

Modified sentence regarding post screening and subgroups for clarification. Old version was probably a relict from DSCQS version.

Change accepted with no objections

Difference Scores:

MP: Diff before or after averaging over subjects?

DH: In DSCQS per subject.

Discussion on what a "condition" is. Average over all files of same HRC? Yes! New term PVS: Processed Video Sequence, one file (SRC x HRC).

Need for definition section in test plan was mentioned by AW.

DMOS = Ref – processed

Agreed

<<Lunch break>>

13:45

Introduction of Jae-Seung KIM and Dong-Hwan Kim, from Samsung Electronics CO Ltd., Digital Media R&D center. Main activity is algorithms and chips for enhancement of pictures on flat TV panels.

Subjective test results data format

Presentation of document from M. Pinson on the format specifying the format of subjective tests results. Using Excel spreadsheet. ("Subject data, temporary document.doc"). Each score of a viewer is a row, containing all the necessary information to identify Lab, subject, etc...

This text has been inserted in section 2.2.1 "Results Data Format" and deleted 2.2.2 "Subject Information Data Format".

Replaced "enable automatic generation of subjective data" by "facilitate data analysis".

A sample spreadsheet should be added.

Added text that clarify that any NR model will bevaluated based on the absolute rating of each PVS (SRCxSRC).

➔ This will be revisited, no agreement yet.

Return to Section 2.2. Data format revised and accepted.

Section 3: test laboratories and schedule

Modified last sentence since it was not practical. Labs now have to report the specs of the test environment they plan to use to the ILGs.

➔ On the MMTEST reflector PC will provide an example report of specs of environment and display to be agreed upon at the next meeting.

Section 3.1

Name of paragraph spelled out now

DSCQS changed to ACR-HRR

INTEL is an ILG lab now

Section 3.2

Added Swissqual and Psytechnics as labs

Details are still to be worked out. A method similar to the selection of P.563 was discussed. Several opinions were stated, but action was referred to working group. Main problem is weighting of "known" versus "unknown" databases.

➔ A proponents working group was initiated (NTIA, BT, Swissqual, Yonsei, Psytechnics, NTT, Genista (? Not present), OPTICOM, KDDI, PB).

➔ CS will try to find ITU contribution describing method applied for P.563

➔ Other proposals are encouraged

DH: Any new proponents?

PB: TDF indicated their interest in being a proponent too.

??: Samsung is not yet sure whether they want to be a proponent or an ILG

OS: KDDI might also become a proponent.

<<Coffee break>>

Presentation of document on LCD response time by KB

Different types of LCDs (TN, S-IPS, PVA or VA) have different response times

Depending on how the time is measured, different results will be achieved. It makes a big difference of a shift between black and white or grey scales (longer) is measured.

Upper chart show switch time from dark to bright, the lower ones vice versa. Each graph in a chart represents a transition to a different grey levels.

Short discussion on converting files to fixed bitrate before viewing to avoid frames shorter than the switch time of the monitor.

DH: Video sequences must fit to the display and the display must be chosen accordingly.

Test Schedule

AW modified some dates in the test schedule which had to be changed for obvious reasons (completen date of the test plan etc.)

Editorial node was added to state that the model exchange related things will be revised by the proponents working group.

Section 4. Sequence Processing and Data Formats

“Three subjective tests …” was replaced by “separate subjective tests…”.

“simulated mobile environment” was removed.

Significant editorial changes

Rec.601 @ 720x486 pixels was added as an option.

VB: Use a mask to cover unused parts of screen.

Provision for aspect ratio correction @ Rec.601 was included in this chapter.

QH: It can be specified in the AVI header to use exactly 1:1 aspect ratio to avoid scaling by the player.

Long discussion on how players will behave in terms of scaling the Rec.601 format. As it seems nobody really knows it.

QH never noticed any difficulties with rescaling if the headers were properly set.

All this will have to be revisited -> Forum!

AW: want to take some decisions.

It was discussed using 4CIF (704x576) or VGA (640x480) instead of Rec.601. The conversion from Rec.601 sources has however to be specified.

➔ It was agreed to change the current test plan from Rec.601 to anything else. This is not yet implemented in the Test plan.

CS: VGA is the smallest of Rec.601/4CIF/VGA. It can be easily cropped from the larger format without scaling effects.

➔ Voting 2:2, many participants were undecided. Discussion was deferred over to the forum and the decision was postponed.

Section 4.?.

Up-/downsampling was related to player.

Section 4.1. Sequence Processing

Exception for animation was removed

Rec. 601 was corrected to VGA/4CIF

MP: Keeping audio in here means that we will have to take care of it when defining the file format.

The presence of audio in the AVI file is depending on the file format which is to be defined more detailed.

CL: It is important to perform deinterlacing and scaling at the same time since both are interpolation operations and combining them allows for better quality. However a method for this has to be found.

Source material must be usable by VQEG MM-proponents and ILG members for testing (NDAs may be required).

Section 4.2. Test Materials

Updated and clarified table a bit

MP: In Rome it was discussed to use SRCs from low quality cameras (e.g. mobile phones) in uncompressed format.

This issue is considered to be ongoing.

Return to Section 2.1.

➔ First sentence shall be changed since not practical. – Accepted with majority.

➔ Paragraph now refers to secret material only.

➔ Detailed procedure of material selection will depend on procedure defined by proponents working group.

Added definition: Secret means a selection out of a large pool. Unknown means no proponents knows the SRC or HRC.

Section 4.3. HRC

Added live network conditions

Section 4.3.1 Bitrates

Discussion to raise the bitrates for PC2 to 4Mbit/s

➔ Proposal accepted after vote (7:3)

Section 4.3.2 Transmission Errors

Presentation of Ericsson's email by KB.

Contents of emails was shifted to an annex as an editors note. To be referred to during the next update of the test plan.

New Section 4.3.3. Live Network Conditions

This shall be filled with the input from Ericsson which is now in the Annex

Section 10 Annex Optimum Mean Square Quantisation Method

Deleted after agreement , seemed to be a relic from TV test and does not apply anymore

Annex 1 Instructions for subjects

Deleted, does not apply to ACR

Will be replaced with new version.

<<Stop>>

Friday, Oct 22nd.

Agenda:

- continue review of test plan
- Write liaisons to ITU.

PC2 image size (4CIF vs. VGA)

Discussion on image size for PC2 profile. 601 is the main source to generate SRCs. Main issues are 1) conversion to lower CIF and QCIF resolutions, 2) pixel aspect ratio conversion to square pixels, 3) black borders.

- 4CIF (704x576) is easy to downconvert to CIF and QCIF, except for 525 601 format.
- CL: proposal to resize vertically to compensate for and crop to generate VGA size from 601.

➔ Vote: 3 for 4CIF, 7 for VGA. VGA chosen.

continuing review of test plan

Review of section "4.3.2. Transmission Errors" and "4.3.3. Frame Freezing and Frame Skipping"

There are a number of error conditions which are not integrated in the test plan. Proposal is to create an ad hoc group on this topic to define transmission errors test conditions, including live network transmission error conditions, that would then release its conclusions on the reflector.

➔ New transmission errors / live network errors group is created, jointly lead by Quan (transmission errors) and Christian (live network conditions). Frame freezing and skipping will also be addressed. A new topic will be created on the MM forum.

Vivaik (intel) would like to see more internet traffic oriented HRCs, in addition to mobile oriented HRCs. Will participate to the Working Group on this topic.

Review of section "4.3.5. Frame rates"

Discussion on minimum frame rates, including for PDA. 5fps for PC1 / PC2, 2.5fps for the PDA mode. Need to address frame rates effectively used in the industry (e.g. 3G mobiles).

➔ Test plan changed: added 8 and 12.5 fps for all PDA PC1 and PC2 modes.

➔ Working group to clarify aspects related to frame rate, effective frame rate and refresh rate, etc... Lead by Arthur W and David H.

Discussion on the statement that the display rate from the player should match the refresh rate of the monitor (at least in CRT case).

➔ This point needs definitions to clarify: source frame rate, player frame rate, monitor refresh rate

<<Coffee break>>

Next meeting date

February or April (3:6).

Dates to be avoided: 6Q meeting proposed dates are april 8-12th. 3GPP conference is on Feb. 14th-17th.

Review of section "6. Objective Quality model evaluation criteria"

Section has been untouched since 3 years, when the subjective evaluation technique had not been chosen. Removal of aspects mainly related to SSCQE.

- Removal of sentence stating that the same techniques used in Phase II will be applied as SSCQE will be used.
- Removal of paragraph, table and figure related to objective quality model evaluation (unclear).
- Removal of table listing the evaluation metrics.

Added a note : the mapping function has to be further discussed

Review of section "6.2 Evaluation metrics"

Discussion on the way how to modify this section. Considering that this section had not been formally agreed and simply pasted from preceding test plan; t is under revision and modifications do not need to be approved by 2/3 majority.

Considering Ericsson's input on the topic (Mmevalstatistics_proposal.doc), as well as emails from Psytechnics and Verizon:

- Metric 2 retained (RMSE). Removal of Metric 1 (Weighted RMSE), metric 2 becomes Metric 1.

<<Lunch>>

- Discussion on relevance of pearson and spearman correlation coefficients. Agreed to remove the spearman correlation coefficient.
- Agreed to remove the 6.2.x headlines describing the relevance of each metric, and to move this topic into the description of each Metric.
- Outlier ratio is kept.
- Discussion on the relevance of Kurtosis in MM. Removed. Philip C stresses on the need to have a measure of distribution normality. Ericsson proposes an histogram with 0.25 MOS bins.
- Kappa removed

This leads to 3 evaluation metrics (with 95% confidence interval), plus 1 significance test, e.g. F-test.

There is a need to clarify at the next meeting the way how to evaluate the global performace of the model by aggregating the performance on individual experiments.

Actions list

Need for definition section in test plan was mentioned by AW. 10DEC04
A sample spreadsheet describing subjective test results format should be added. MPinson 5NOV04
On the MMTEST reflector PHILC will provide an example report of specs of environment and display to be agreed upon at the next meeting 12NOV04.
A proponents working group was initiated (NTIA, BT, Swissqual, Yonsei, Psytechnics, NTT, Genista (? Not present), OPTICOM, KDDI, PB).Christian new reflector 5NOV04
ChrisS will try to find ITU contribution describing method applied for P.563 5NOV04
It was discussed using 4CIF (704x576) or VGA (640x480) instead of Rec.601. The conversion from Rec.601 sources has however to be specified. Discuss on MMFORUM: Decision at audio call DEC04
a method for to perform deinterlacing and scaling has to be found. . Discuss on MMFORUM: Decision at audio call DEC04
Audio call for MM, PhilC to provide bridge around Middle of December(e.g. 8 or 9 Dec)
New transmission errors / live network errors group is created, jointly lead by Quan (transmission errors) and Chritian (live network conditions). Frame freezing and skipping will also be addressed. A new topic will be created on the MM forum. New Topic on Forum done: Email to be sent. To reflector by 5NOV04
Working group to clarify aspects related to frame rate, effective frame rate and refresh rate, etc... point needs definitions to clarify: source frame rate, player frame rate, monitor refresh rate Lead by Arthur W and David H. with definitions group to be done by 10Dec.
the mapping function between MOS and MOSp has to be further discussed DATA Analysis discussion to be Kjell due by 10Dec04
There is a need to clarify at the next meeting the way how to evaluate the global performace of the model by aggregating the performance on individual experiments. Led by Chulhee Lee Test Schedule subgroup led by David Hands by 7DEC04 Calibration section to be drafted by M.Pinson by 7DEC04.