IEICE TRANS. COMMUN., VOL.E97–B, NO.1 JANUARY 2014

PAPER

Performance Comparisons of Subjective Quality Assessment Methods for Video

Toshiko TOMINAGA[†a)], **Masataka MASUDA**[†], **Jun OKAMOTO**[†], **Akira TAKAHASHI**[†], *and* **Takanori HAYASHI**[†], *Members*

SUMMARY Many subjective assessment methods for video quality are provided by ITU-T and ITU-R recommendations, but the differences among these methods have not been sufficiently studied. We compare five subjective assessment methods using four quantitative performance indices for both HD and QVGA resolution video. We compare the Double-Stimulus Continuous Quality-Scale (DSCQS), Double-Stimulus Impairment Scale (DSIS), Absolute Category Rating method (ACR), and ACR with Hidden Reference (ACR-HR) as common subjective assessment methods for HD and QVGA resolution videos. Furthermore, we added ACR with an 11-grade scale (ACR11) for the HD test and Subjective Assessment of Multimedia Video Quality (SAMVIQ) for the QVGA test for quality scale variations. The performance indices are correlation coefficients, rank correlation coefficients, statistical reliability, and assessment time. For statistical reliability, we propose a performance index for comparing different quality scale tests. The results of the performance comparison showed that the correlation coefficients and rank correlation coefficients of the mean opinion scores between pairs of methods were high for both HD and QVGA tests. As for statistical reliability provided by the proposed index, DSIS of HD and ACR of QVGA outperformed the other methods. Moreover, ACR, ACR-HR, and ACR11 were the most efficient subjective quality assessment methods from the viewpoint of assessment time.

key words: *video quality, subjective assessment method, performance comparison, HD, QVGA, H.264*

1. Introduction

Internet Protocol Television (IPTV) and mobile video streaming services are expanding. To provide these services with appropriate quality, service quality parameters based on a user's quality of experience (QoE) is important. Subjective quality assessments are still the most accurate way to evaluate QoE. Therefore, service quality parameters, such as encoding video bitrate target, video coding parameter optimization, and network quality target, are determined through subjective quality assessments. Several subjective assessment methods suitable for video application are recommended by ITU-T and ITU-R.

The Double-Stimulus Continuous Quality-Scale (DSCQS) method [1] is a widely used subjective assessment method for video quality in broadcasting. The Double-Stimulus Impairment Scale (DSIS) method [1], [2] (also called the DCR or EBU method) and the Absolute Category Rating method (ACR) with a five-grade scale [1], [2] are also used. In addition, ACR with Hidden Reference (ACR-HR) [2] was recently developed for equalizing the effect diverse source-video qualities have on subjective assessment values. The ACR method with an 11-grade scale (ACR11) is listed as an additional scale that has higher distinguishing power [2]. Moreover, the Subjective Assessment of Multimedia Video Quality (SAMVIQ) method[3] was developed as a quality assessment method suited for a variety of video formats and multimedia environments, such as video streaming for personal computers and mobile terminals. In SAMVIQ, evaluation videos with explicit and hidden reference are used.

These subjective quality assessment methods have specific characteristics and can be classified as follows. The first classification of the subjective assessment method is how to give the stimulus; single stimulus (absolute quality) or a double stimulus (relative quality). The DSCQS, DSIS, and SAMVIQ methods are double stimulus methods and show a viewer explicit reference videos. The ACR, ACR-HR, and ACR11 methods are single stimulus methods.

The second classification is the score calculation method. The DSIS, ACR, ACR11, and SAMVIQ methods use the score directly. On the other hand, DSCQS and ACR-HR indirectly calculate the score by processing both the scores of evaluation video and (hidden) reference video.

The third classification is a quality scale resolution such as discrete or continuous quality scale. The ACR, ACR-HR, and DSIS methods have discrete quality scales, such as five categories with five quality scales, and ACR11 has five categories with 11 grade quality scale. The DSCQS and SAMVIQ methods have continuous scales such as five categories with 100 quality scales.

Another classification method is the flexibility of evaluation. In SAMVIQ, a viewer operates a PC and can freely choose the evaluation video and watch the explicit reference whenever he/she wants. On the other hand, in almost all other evaluation methods, the presentation order of the video sequence is fixed and not controled by the viewer.

There are many subjective assessment methods, so it is not clear which method is the most suitable for evaluating QoE characteristics. A performance comparison of subjective quality assessment methods was conducted to show that a new subjective assessment method, such as ACR-HR or SAMVIQ, had performed better than the conventional ones [4]–[11]. Moreover, Winkler and Quan conducted performance comparisons with additional quality scales [12], [13], but the number of viewers and variety of video se-

Manuscript received January 4, 2013.
Manuscript revised September 5, 2013.

[†]The authors are with NTT Network Technology Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: toshiko.tominaga@lab.ntt.co.jp

DOI: 10.1587/transcom.E97.B.66

quences were different for each experiment [12] and only absolute category ratings methods were used [13]. Performance comparison requires various viewpoints. Correlation coefficients and rank order correlation coefficients are often used as performance indices when comparing subjective assessment methods [4], [6]–[11], [13], [14].

To understand the stability and distinction of the mean opinion score (MOS), it is also necessary to consider the statistical reliability, which is a confidence interval of the MOS, as a quality assessment index. Statistical reliability cannot be compared directly when the rating scale is different. Therefore, a unique measure is required to compare subjective assessment methods that have different rating scales. Two conventional methods for such a unique measure have been proposed: the Score Transformation method [4]–[6], [9], [11], [13] and the Normalizing Score with a Rating-scale Range method [12]. If the quality measure is the same from "Bad" to "Excellent," the score distributions are not always the same with different rating scales such as those with 5 grades, 11 grades, or a continuous scale. For a unique measure of statistical reliability, there was no sufficient consideration in terms of score distributions. Therefore, we evaluated the total statistical reliability using the mean of 95% confidence interval (MCI) through the entire MOS range (from low to high MOS) [14].

Furthermore, subjective assessment methods require much money and time. When the same performance is being assessed, a shorter assessment time is more efficient. Experimental efficiency has not been included in any performance comparison index [14].

We had performed performance comparison for QVGA video [14], but clarifying the difference in assessment characteristics for different video formats are also needed. Péchard et al. used a variety of video formats, QVGA, CIF, VGA, and HD for ACR and SAMVIQ methods [11]. Winkler used both CIF and SD formats for five different quality scales, but did not use the same video materials [12].

We clarified video quality characteristics by different subjective assessment methods using the same comparison performance indices. We used four performance comparison indices: correlation coefficients, rank correlation coefficients, statistical reliability, and assessment time. As the statistical reliability index, we propose a 95% confidence interval (CI) normalized by the MOS range for different scales. After verifying the validity of the above reliability index, we conducted a statistical reliability comparison. For experimental efficiency, we used the assessment time of one video sequence. We compared five subjective assessment methods for the same resolution videos. The HD format was used for IPTV services and QVGA format was used for mobile video streaming services. We also focused on H.264 videos. The DSCQS, DSIS, ACR, and ACR-HR methods were used as common assessment methods for HD and QVGA tests. We also added ACR11 for HD and SAMVIQ for QVGA for quality scale variations. The SAMVIQ software had a performance limit, so we could not conduct a SAMVIQ experiment for HD video. In the experiment for HD video, we used ACR11 instead of SAMVIQ, which has more than five grades on the quality scale.

The remainder of this paper is structured as follows. Section 2 presents the subjective assessment test methods for video quality. Performance indices for comparison are described in Sect. 3. The HD and QVGA video assessment tests and their performance comparison results are described in Sects. 4 and 5. A conclusion is given in Sect. 6.

2. Subjective Assessment Methods

This section describes the subjective assessment methods used for performance comparison.

2.1 DSCQS Method

This method requires the assessment of two versions of each test sequence. One of each pair of test sequences is unimpaired, whereas the other presentation might or might not contain an impairment. The unimpaired sequence is included to serve as a reference, but the viewers are not told which is the reference sequence. In the series of tests, the position of the reference sequence is changed in pseudo random order (Fig. 1).

- Quality scale: The viewers are simply asked to assess the overall video quality of each presentation by inserting a mark on a vertical scale. The vertical scales are printed in pairs to accommodate the double presentation of each test sequence. The scales provide a continuous rating system to avoid quantizing errors, but they are divided into five equal lengths that correspond to a five-point quality scale (Bad, Poor, Fair, Good, Excellent). The assessment pairs (reference and test sequences) for each test condition are converted from

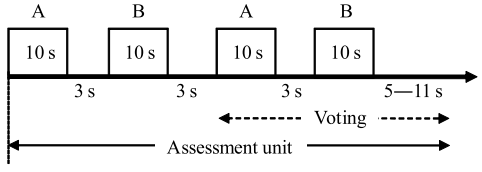

Fig. 1 DSCQS method.

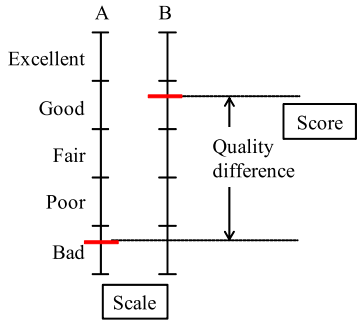

Fig. 2 DSCQS quality scale.

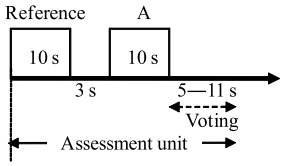

Fig. 3 DSIS method.

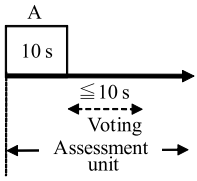

Fig. 4 ACR and ACR11 methods.

measurements of length on the score sheet to normalized scores ranging from 0 to 100 (Fig. 2).

- Quality expression: Quality is expressed by the differential viewer scores for the processed video sequence ($DV(PVS)$). $DV(PVS)$ can be calculated using the difference between the viewer scores for PVS ($V(PVS)$) and the viewer scores for the reference ($V(REF)$) as follows;

$$DV(PVS) = | V(PVS) - V(REF) | . \quad (1)$$

2.2 DSIS Method

The DSIS method implies that the video sequences are presented in pairs: the first sequence presented in each pair is always the reference sequence, while the second sequence is the test sequence (Fig. 3).

- Quality scale: Each viewer selects a category from a discrete scale of five categories (1: Very annoying, 2: Annoying, 3: Slightly annoying, 4: Perceptible but not annoying, 5: Imperceptible).
- Quality expression: Quality is expressed by $V(PVS)$.

2.3 ACR and ACR-HR Method

The ACR method is a category judgment method in which the test sequences are presented one at a time (Fig. 4).

- Quality scale: Each viewer selects a category from a discrete scale of five categories (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).
- ACR Quality expression: Quality is expressed by $V(PVS)$.
- ACR-HR Quality expression: Quality is expressed by the DV scores as follows using PVS and hidden reference (HREF).

$$DV(PVS) = V(PVS) - V(HREF) + 5. \quad (2)$$

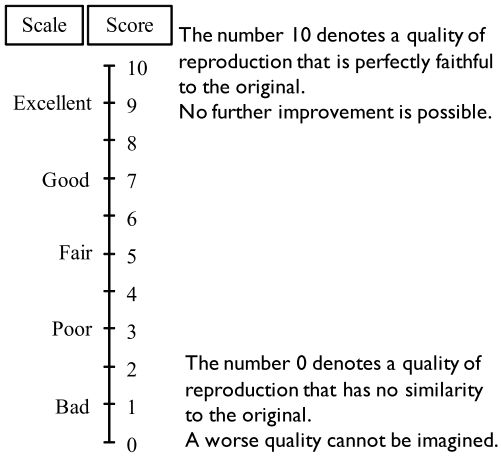

Fig. 5 ACR11 quality scale.

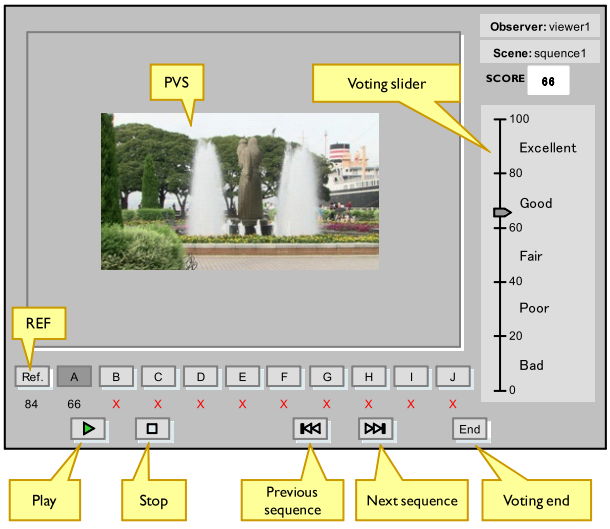

Fig. 6 SAMVIQ evaluation display image.

2.4 ACR11 Method

The ACR11 method is a category judgment method in which the test sequences are presented one at a time (Fig. 4).

- Quality scale: Each viewer selects a grade from a discrete scale of 11 grades in 5 categories (Fig. 5).
- ACR11 Quality expression: Quality is expressed by $V(PVS)$.

2.5 SAMVIQ Method

In this method, the viewer is given access to several versions of a sequence. When all versions have been rated by the viewer, the following sequence content can then be accessed. The different versions are randomly selectable by the viewer through a computer graphic interface. The viewer can stop, review, and modify the score of each version of a sequence as desired. Quality evaluation is carried out scene

by scene (see Fig. 6 [10]), including an explicit reference and an HREF.

- Quality scale: This method uses a continuous quality scale to provide a measurement of the intrinsic quality of video sequences. Each viewer moves a slider on a continuous scale from 0 to 100 annotated by 5 quality items linearly arranged (Bad, Poor, Fair, Good, Excellent).
- SAMVIQ Quality expression: Quality is expressed by $V(PVS)$.

3. Performance Indices for Comparison

We used four performance indices to compare the assessment methods in terms of sensitivity, reliability, and efficiency. The correlation and rank correlation coefficients are the indices for sensitivity. Statistical reliability is an index for reliability and the assessment time per condition is an index for efficiency. For statistical reliability, we use our performance index. The details of the four comparison indices are described below.

(1) Correlation coefficient
The Pearson product-moment correlation coefficient (R) is an index that shows the correspondence-related strength of the MOS for every condition. A high correlation coefficient shows high correspondence between two methods. The formula of R between assessment methods A and B is as follows.

$$R = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \quad (3)$$

where
x_i represents the MOS for the ith test condition in assessment method A,
y_i represents the MOS for the ith test condition in assessment method B, and
$\overline{x}$ represents the average value of MOS for all the test conditions in assessment method A,
$\overline{y}$ represents the average value of MOS for all the test conditions in assessment method B, and
n represents the number of test conditions

(2) Rank correlation coefficient
Spearman's rank correlation coefficient ($rankR$) is a measure that shows the correlation degree of rank for MOSs. A rank correlation coefficient near 1 shows a high rank relation. The formula of $rankR$ between assessment methods A and B is as follows.

$$rankR = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n^3 - n}, \quad (4)$$

where
x_i represents the MOS order for the same test condition in assessment method A,
y_i represents the MOS order for the same test condition in assessment method B, and
n represents the number of test conditions

(3) Statistical reliability
To compare the statistical reliability of the MOS, we used our 95% CI since the CI values are based on the standard deviation of the MOS. For comparing different quality scales, there are two conventional methods; the Score Transformation method [4]–[6], [9], [11], [13] and the Normalizing Score with a Rating-scale Range method [12]. Even if we used the same PVS, each method has a different range of MOSs. This is because each method has its own framework of the test. Therefore, we propose using our normalized 95% CI by using the MOS range and considering the framework of the test.

(4) Assessment time
Subjective quality assessment requires a large amount of time and cost. The assessment time consists of the video presentation time and voting time per condition. A shorter assessment time is preferable if it produces the same assessment accuracy.

4. Comparison of Subjective Assessment Methods for HD Video

This section describes performance comparison results of the subjective assessment methods for HD video. First, we describe the test conditions used in the five subjective test methods i.e., DSCQS, DSIS, ACR-HR, ACR, and ACR11. Then, we discuss the comparison results of these methods using four indices.

4.1 Subjective Assessment Test Conditions

We conducted four experiments with the four subjective assessment methods (DSCQS, DSIS, ACR, and ACR11). The experiment with ACR-HR is unnecessary because it is a method of calculating MOSs using the experimental results from ACR. The SAMVIQ software had a performance limit, so we could not conduct a SAMVIQ experiment for HD video. In the experiment for HD video, we used ACR11 instead of SAMVIQ, which has more than five grades on the quality scale.

Eight videos in ITU-R recommendation BT.1210-3 [15] were used as the source material and are listed in Table 1. The video quality degradation factor was the encoding bit rate. Video sources were encoded by H.264. The features of H.264 coding distortions are block noise and jerkiness, which are the same as those of other encoding methods. We used 56 PVSs, which consisted of 8 source videos and 48 encoded videos. Details of the video quality conditions are listed in Table 2. Table 3 summarizes the

Table 1 Video sources.

No. [15]	Title
5	Boy and toys
7	European market
10	Street car
12	Harbour scene
16	Whale show
23	Green leaves
25	Japanese room
34	Ice hockey

Table 2 Video quality conditions for HD.

Video length	10 sec
Codec	H.264 High Profile (Tandberg EN5990)
Video format	HD (1440x1080)
Bit rate: BR (Mbps)	3.0, 4.3, 5.7, 6.9, 9.6, 15
Frame rate (fps)	30

Table 3 Viewing conditions for HD.

Monitor	32-inch CRT monitor SONY BVM-D32E1WJ
Display resolution	1920x1080
Viewing distance	3H (H: picture hight); about 110 cm
Room illuminance	about 20 lx

specifications for the viewing conditions [1]. The participants watched a single HD monitor. Forty-eight non-expert participants (24 males and 24 females aged 20–39) joined in the experiments. Each participant performed all assessment methods in one day. The four methods were executed in a different order. We used a total of six randomized order patterns. We also changed the presentation orders of the PVS for each test. The experiment time of each method was measured.

4.2 Performance Comparison Results

4.2.1 Correlation Coefficient

First, we evaluated the score distribution. The score distributions of the 48 PVSs and 8 references for all 48 participants for each method are shown in Fig. 7. For DSCQS, the step size of rating was set to 10 to obtain the distribution characteristics. Individual scores were distributed over the entire quality scale in each assessment method. Next, we evaluated the MOS sensitivity. The MOS relationships between DSCQS and the other four methods are shown in Fig. 8. We calculated R to clarify the relationships and characteristics.

Table 4 lists the Rs between pairs of methods for 48 PVSs. All the Rs for DSCQS were negative. This is because smaller values indicate higher quality in DSCQS, as opposed to the other four methods, where smaller values indicate lower quality. The average absolute values of R for each method are listed in the bottom row of Table 4. From the t-test results at a 5% significance level, there were no significant differences in the average R for each method. This means that each method showed the same assessment performance in R.

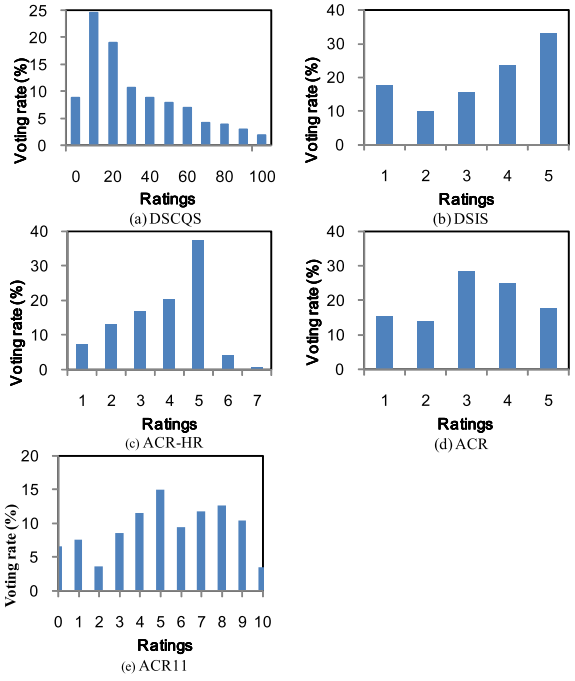

Fig. 7 Distribution of ratings for HD.

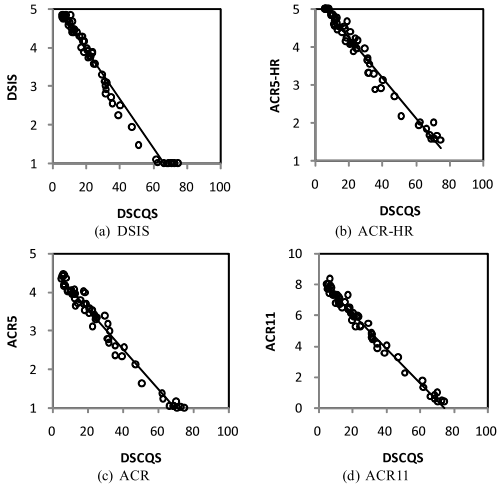

Fig. 8 Correlation between DSCQS and other methods for HD.

Table 4 R between pairs of methods for HD.

	DSCQS	DSIS	ACR-HR	ACR	ACR11
DSCQS	-	−0.99	−0.99	−0.99	−0.99
DSIS	−0.99	-	0.99	0.99	0.99
ACR-HR	−0.99	0.99	-	1.00	0.99
ACR	−0.99	0.99	1.00	-	0.99
ACR11	−0.99	0.99	0.99	0.99	-
Average*	0.99	0.99	0.99	0.99	0.99

*: Average to absolute value of correlation coefficients

Table 5 Rank R between pairs of methods for HD.

	DSCQS	DSIS	ACR-HR	ACR	ACR11
DSCQS	-	0.99	0.98	0.98	0.98
DSIS	0.99	-	0.98	0.97	0.98
ACR-HR	0.98	0.98	-	0.99	0.98
ACR	0.98	0.97	0.99	-	0.98
ACR11	0.98	0.98	0.98	0.98	-
Average	0.98	0.98	0.98	0.98	0.98

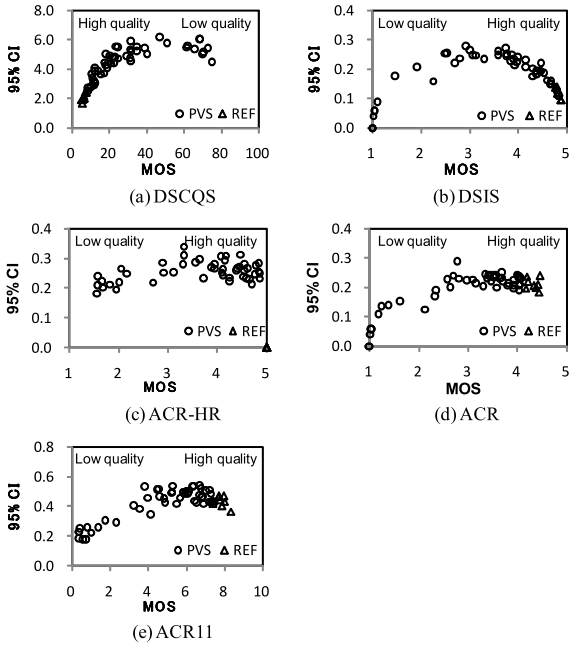

Fig. 9 MOS and 95% CI for HD.

4.2.2 Rank Correlation Coefficient

We calculated *rankR* to clarify the characteristics of the MOS ranking. Table 5 lists the *rankR*s between pairs of methods for 48 PVSs. The averaged *rankR*s for each method are listed in the bottom row in Table 5. From the result of t-test at a 5% significance level, there were no significant differences in the average *rankR* for each method. This means that each method showed the same assessment performance in *rankR*.

4.2.3 Statistical Reliability

The MOS and its 95% CI for PVSs and REFs are shown in Fig. 9. The X-axis shows the MOS for each quality scale and the Y-axis shows its 95% CI. Generally, characteristics of CI distribution of the MOS have small CI values at both ends of the MOS range and have large CI values at the middle of the MOS range. To compare statistical reliability, we believe it is important to understand the difference in the 95% CI characteristics. The ACR and ACR11 methods have a smaller 95% CI in the low quality section than that in the high quality section. On the contrary, DSCQS has a smaller 95% CI in the high quality section than that in the low quality section. The DSIS method has a small 95% CI in both high and low quality sections. As mentioned above, each method has a different characteristic of the 95% CI distribution. Score statistics are given in Table 6. We focused on the quality scale use rate. That is, even if the same PVSs are used, the scale use rate changes with the method. This is because each method has a different framework. Therefore, we use our normalized 95% CI by using the MOS range as a statistical reliability index. We evaluated the proposed index [14] as well as two conventional indices, which are the Score Transformation method [4]–[6], [9], [11], [13] and the Normalizing Score with a Rating-scale Range method [12]. They are defined as follows.

Table 6 Score statistics for HD.

	DSCQS	DSIS	ACR-HR	ACR	ACR11
Max. scale	100	5	5	5	10
Min. scale	0	1	1	1	0
Scale range	100	4	4	4	10
Max. MOS(PVS)	74.71	4.83	4.85	4.06	7.44
Min. MOS(PVS)	8.73	1.00	1.56	1.00	0.40
MOS range(PVS)	65.98	3.88	3.29	3.06	7.04
Scale use rate(%)	66%	96%	82%	77%	70%

Table 7 MCIs and those of three indices for HD.

MCI	DSCQS	DSIS	ACR-HR	ACR	ACR11
Original	4.581	0.178	0.256	0.190	0.421
Index A	0.069	0.046	0.078	0.062	0.060
Index B	0.183	0.178	0.256	0.190	0.169
Index C	0.046	0.045	0.064	0.047	0.042

(Index A; proposed index) Normalized 95% CI by using the MOS range ($95\%CI_{MOS\ range}$),

$$95\%CI_{MOS\ range} = \frac{95\%CI}{MOS\ range}. \tag{5}$$

(Index B) Score transformation from DSCQS ($Score_{0-100}$), SAMVIQ ($Score_{0-100}$) and ACR11 ($Score_{0-10}$) on a five-grade scale ($Score_{1-5}$),

$$Score_{1-5} = \frac{Score_{0-100}}{25} + 1, \quad for\ DSCQS\ and\ SAMVIQ, \tag{6}$$

$$Score_{1-5} = \frac{Score_{0-10}}{2.5} + 1, for\ ACR11, \tag{7}$$

then calculate 95% CI.

(Index C) Normalized 95% CI by using a Rating-scale Range ($95\%CI_{Scale\ range}$),

$$95\%CI_{Scale\ range} = \frac{95\%CI}{Scale\ range}. \tag{8}$$

We compared three statistical reliability indices of the MCI. Table 7 lists the MCIs and those of three statistical reliability indices for each method. We used only PVSs for index comparison. This is because $DV(REF)$ of ACR-HR is 5, so the 95% CI is always 0. We then conducted t-tests at a 5% significance level between the MCI of DSCQS

Table 8 T-test results (p-value) for MCI of three indices between DSCQS and other methods for HD.

p-value	DSIS	ACR-HR	ACR	ACR11
Index A	0.000^{**}	0.002^{**}	$0.105^{n.s.}$	0.011^{*}
Index B	$0.714^{n.s.}$	0.000^{**}	$0.612^{n.s.}$	$0.145^{n.s.}$
Index C	$0.714^{n.s.}$	0.000^{**}	$0.612^{n.s.}$	$0.145^{n.s.}$

n.s.: not significant, *: $p < 0.05$, **: $p < 0.01$

Table 9 Assessment time per condition for each method for HD.

	DSCQS	DSIS	ACR-HR	ACR	ACR11
Assessment time(sec)	54	28	15	15	15

and that of the other methods and list their p-values in Table 8. The DSCQS method is used for user requirements for TV systems [16]. Therefore, we used this as a combination reference in the t-test. In indices B and C, which are conventional indices, there were no significant differences between DSCQS and DSIS and ACR and ACR11. The ACR-HR method had significantly larger statistical reliability than DSCQS. On the other hand, index A, which takes into account the framework of the test, showed that DSIS and ACR11 had a smaller MCI than DSCQS. The ACR method had no significant differences to DSCQS in terms of the MCI. The ACR-HR method had a larger MCI than DSCQS. From the results of index A, DSIS had the smallest MCI of the five methods. In indices B and C, which are conventional indices, there were no significant differences between DSCQS and DSIS and ACR and ACR11. The ACR-HR method had significantly larger statistical reliability than DSCQS. As a reference, we confirmed MCI comparison of the three MOS range sections; lower 25% (low MOS range), middle 50% (middle MOS range), and upper 25% (high MOS range). Though the number of conditions in each MOS range section were not the same for each assessment method, the three MOS range sections' results were the same as those of the entire MOS range.

4.2.4 Assessment Time

We compared the assessment time between methods from the viewpoint of experimental efficiency. The assessment time consisted of the time to view each PVS and the evaluation time. The maximum evaluation time was 5 sec. In the DSCQS and DSIS tests, the interval between a PVS and the reference PVS was 3 sec. Table 9 lists the assessment time per condition for each method. The ACR, ACR-HR, and ACR11 methods had shorter assessment times than the other methods.

4.2.5 Performance Comparison Results for HD

The performance comparison results for HD showed that the correlation and rank correlation coefficients of the MOSs between pairs of methods were high. For statistical reliability, index A showed that DSIS had the smallest MCI and both indices B and C had equivalent MCIs to those of DSCQS, DSIS, ACR, and ACR11. The ACR, ACR-HR, and ACR11 methods are efficient from the viewpoint of assessment time.

Table 10 Video quality conditions for QVGA.

Video length	10 sec
Codec	H.264 Main Profile (Quick Time 7 Pro for Win. Ver. 7.6.4)
Format	QVGA (320x240)
Bit rate: BR (kbps)	128, 256, 384, 704, 1500
Frame rate (fps)	30

Table 11 Viewing conditions for QVGA.

Monitor	17-inch LCD monitor EIZO FLEXSCAN M170
Display resolution	1280x1024
Viewing distance	8H (H: picture hight); about 50 cm
Room illuminance	about 20 lx

5. Comparison of Subjective Assessment Methods for QVGA Video

This section describes the performance comparison results of the subjective assessment methods for QVGA video.

5.1 Subjective Assessment Test Conditions

We conducted four experiments (DSCQS, DSIS, ACR, and SAMVIQ). The ACR-HR methods' MOSs were calculated using the experimental results from ACR. The eight video sources for the HD test were also used in the QVGA test. The video sources were resized to QVGA (320×240 pixels) and encoded by H.264. The video quality degradation factor was the encoding bit rate. We used 8 REFs and 40 PVSs. Details of the video quality conditions are listed in Table 10.

A 4-inch video window was displayed on a 17-inch liquid crystal display (1280×1024 pixels). The viewing conditions are listed in Table 11 [2]. Forty-eight non-expert participants (24 males and 24 females aged 20–39) participated in the experiments. Each participant did all assessments in one day. The executing order of the four methods and the PVS randomization in each test were the same as in the HD test. The experiment time of each method was measured. In the SAMVIQ test, a viewer can repeat the video sequence and each viewer's experiment time is different. Therefore, we measured the experiment time per participant.

5.2 Performance Comparison Results

5.2.1 Correlation Coefficient

First, we evaluated the score distribution. The score distributions for 40 PVSs and 8 REFs of the 48 participants for each method are shown in Fig. 10. For the DSCQS method, the step size of rating was set to 10 to obtain the distribution characteristics. Individual scores were distributed over the entire quality scale in each assessment method. The MOS and 95% CI for each PVS are plotted in Fig. 11. The CI values in the QVGA evaluation results were larger than those

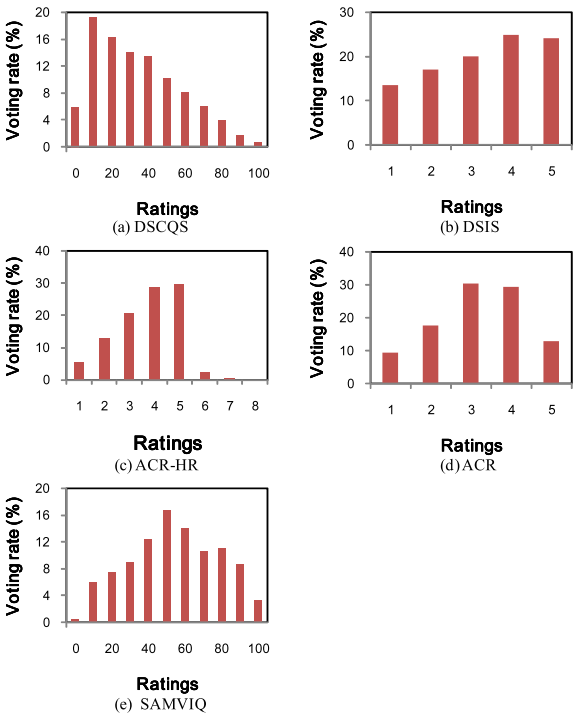

Fig. 10 Distributions of ratings for QVGA.

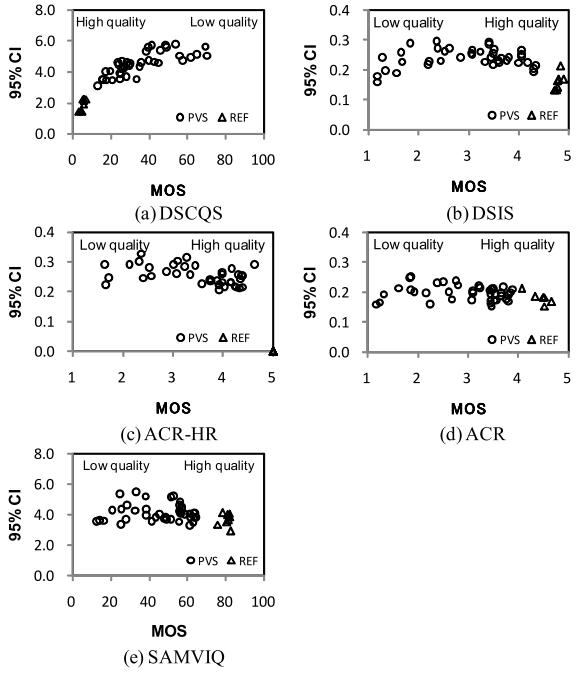

Fig. 11 MOS and 95% CI for QVGA.

in HD evaluation results shown in Fig. 9. QVGA video displays at a high resolution, and long viewing distance can be considered as a cause of large CI values. We plan to investigate the cause of such large CI values in the QVGA evaluation. Score statistics are given in Table 12. The scale use rates in the QVGA test were narrower than those of the HD test (Table 6). The relationships between DSCQS and

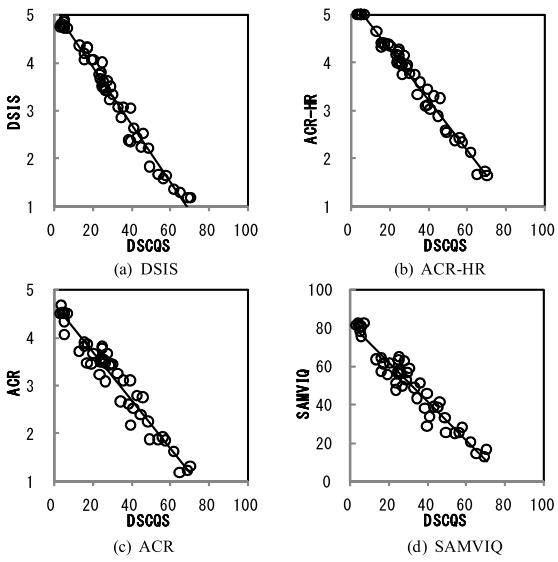

Fig. 12 Correlation between DSCQS and other methods for QVGA.

Table 12 Score statistics for QVGA.

	DSCQS	DSIS	ACR-HR	ACR	SAMVIQ
Max. scale	100	5	5	5	100
Min. scale	0	1	1	1	0
Scale range	100	4	4	4	100
Max. MOS (PVS)	70.46	4.35	4.65	3.90	64.65
Min. MOS (PVS)	13.25	1.19	1.65	1.19	12.85
MOS range (PVS)	57.21	3.17	3.00	2.71	51.79
Scale use rate (%)	57%	79%	75%	68%	52%

Table 13 R between pairs of methods for QVGA.

	DSCQS	DSIS	ACR-HR	ACR	SAMVIQ
DSCQS	-	−0.99	−0.99	−0.97	−0.96
DSIS	−0.99	-	0.99	0.97	0.97
ACR-HR	−0.99	0.99	-	0.98	0.97
ACR	−0.97	0.97	0.98	-	0.99
SAMVIQ	−0.96	0.97	0.97	0.99	-
Average*	0.98	0.98	0.99	0.98	0.98

*: Average to absolute value of correlation coefficients

the other four methods are plotted in Fig. 12. Table 13 lists the *R*s between pairs of methods for 40 PVSs. The averaged absolute values of *R* for each method are listed in the bottom row in Table 13. From the t-test results at a 5% significance level, there were no significant differences in the average *R* for each method. This means that each method shows the same assessment performance in *R*.

5.2.2 Rank Correlation Coefficient

We calculated *rankR* for the QVGA test. Table 14 lists the *rankR*s between pairs of methods for 40 PVSs. The averaged *rankR* for each method are listed in the bottom row of Table 14. From the result of t-test at a 5% significance level, there were no significant differences in the average *rankR* for each method. This means that each method shows the same assessment performance in *rankR*.

Table 14 Rank R between pairs of methods for QVGA.

	DSCQS	DSIS	ACR-HR	ACR	SAMVIQ
DSCQS	-	0.99	0.98	0.94	0.93
DSIS	0.99	-	0.99	0.96	0.95
ACR-HR	0.98	0.99	-	0.95	0.95
ACR	0.94	0.96	0.95	-	0.98
SAMVIQ	0.93	0.95	0.95	0.98	-
Average	0.96	0.97	0.97	0.96	0.96

Table 15 MCI and those of three indices for QVGA.

MCI	DSCQS	DSIS	ACR-HR	ACR	SAMVIQ
Original	4.504	0.239	0.255	0.199	4.130
Index A	0.079	0.076	0.085	0.073	0.080
Index B	0.180	0.239	0.255	0.199	0.165
Index C	0.045	0.060	0.064	0.050	0.041

Table 16 T-test results (p-value) for MCI of three indices between DSCQS and other methods for QVGA.

p-value	DSIS	ACR-HR	ACR	SAMVIQ
Index A	$0.210^{n.s.}$	0.002^{**}	0.023^{*}	$0.649^{n.s.}$
Index B	0.000^{**}	0.000^{**}	0.002^{**}	0.005^{**}
Index C	0.000^{**}	0.000^{**}	0.002^{**}	0.005^{**}

n.s.: not significant, *: $p < 0.05$, **: $p < 0.01$

5.2.3 Statistical Reliability

We compared three statistical reliability indices of the mean 95% CI (MCI) for PVSs. The MCI and those of the three indices for QVGA are listed in Table 15. The p-values of the t-tests at 5% significance level for the MCIs of three indices are listed in Table 16. The DSCQS method is used for user requirements for TV system [16]. Therefore, we used this as a combination reference in the t-test. Indices B and C, which have no framework consideration, showed that SAMVIQ had a smaller MCI than DSCQS. On the other hand, ACR had a smaller MCI than DSCQS by using index A, which has framework consideration. As a result, index A showed that ACR had the smallest MCI and both indices B and C showed that SAMVIQ had the smallest MCI. As a reference, we confirmed MCI comparison of the three MOS range sections; lower 25% (low MOS range), middle 50% (middle MOS range), and upper 25% (high MOS range). Though the number of conditions in each MOS range section were not the same for each assessment method, the three MOS range sections' results were the same as those of the entire MOS range.

5.2.4 Assessment Time

Table 17 lists the assessment time of each method per condition for QVGA. As for assessment time for the SAMVIQ, we calculated the average assessment time per condition per participant because SAMVIQ allows repetitive viewing and voting of the PVS. The result that the assessment time of SAMVIQ is less than DSCQS is the same as that in ITU-R document 6Q/208-E [10]. As a result, ACR and ACR-HR had shorter assessment times than the other methods.

Table 17 Assessment time per condition for QVGA.

	DSCQS	DSIS	ACR-HR	ACR	SAMVIQ
Assessment time(sec)	54	28	15	15	41

5.2.5 Performance Comparison Results for QVGA

The performance comparison results for QVGA showed that the correlation and rank correlation coefficients of the MOSs between pairs of methods were high. For statistical reliability, index A showed that ACR had the smallest MCI and both indices B and C showed that SAMVIQ had the smallest MCI. Moreover, ACR and ACR-HR were the most efficient methods from the viewpoint of assessment time.

6. Conclusion

There are many subjective assessment methods, but it was not clear which method is the most suitable for obtaining QoE characteristics in terms of sensitivity, reliability, and efficiency. We compared five subjective assessment methods using four performance indices for both HD and QVGA resolution video. The performance indices were correlation coefficients, rank correlation coefficients, statistical reliability, and assessment time. Statistical reliability cannot be compared directly when the rating scale is different. Therefore, we proposed a performance index by taking into account the framework of the test.

Performance comparison results showed that the correlation and rank correlation coefficients of the MOSs between pairs of methods were high for both HD and QVGA tests. For statistical reliability, index A showed that DSIS had the smallest MCI and both indices B and C had equivalent MCIs to those of DSCQS, DSIS, ACR, and ACR11 in the HD test. In the QVGA test, index A showed that ACR had the smallest MCI and both indices B and C showed that SAMVIQ had the smallest MCI. Moreover, ACR, ACR-HR, and ACR11 were the most efficient methods from the viewpoint of assessment time. Performance comparison of subjective assessment methods for degraded video quality due to packet losses is for further study.

Acknowledgements

This research was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication of Japan (No. 073103002).

References

[1] ITU-R Rec. BT.500-12, "Methodology for the subjective assessment of the quality of television pictures," Sept. 2009.
[2] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.
[3] ITU-R Rec. BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications," Jan. 2007.

[4] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," Signal Process., vol.77, pp.1–9, Aug. 1999.
[5] F. Speranza, T. Martin, and R. Renaud, "Subjective quality assessment and the effect of context in expert and non-expert viewers," in Proc. SPIE Image Quality Syst. Perform., vol.5294, pp.201–210, San Jose, Jan. 2004.
[6] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," Proc. SPIE Visual Commun. Image Process., Lugano, vol.5150, pp.573–582, July 2003.
[7] C. Lee, H. Choi, E. Lee, S. Lee, and J. Choe, "Comparison of various subjective video quality assessment methods," Proc. SPIE Image Quality Syst. Perform. III, vol.6059, San Jose, Jan. 2006.
[8] ITU-R Document 6Q/131-E, "Technical report: Comparison of DSCQS and ACR," Oct. 2005.
[9] M. Brotherton, Q. Huynh-Thu, D. Hands, and K. Brunnström, "Subjective multimedia quality assessment," IEICE Trans. Fundamentals, vol.E89-A, no.11, pp.2920–2932, Nov. 2006.
[10] ITU-R Document 6Q/208-E, "Report on experiment of new subjective video quality metrics SAMVIQ for mobile video," April 2007.
[11] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm," IMQA 2008, pp.236–241, Sept. 2008.
[12] S. Winkler, "On the properties of subjective ratings in video quality experiments," Proc. Int. Workshop Quality Multimedia Exper. (QoMEX), San Diego, CA, July 2009.
[13] H.-T. Quan, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," IEEE Trans. Broadcast., vol.57, no.1, pp.1–14, March 2011.
[14] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods," QoMEX 2010, pp.82–87, July 2010.
[15] ITU-R Rec. BT.1210-3, "Test materials to be used in subjective assessment," Feb. 2004.
[16] ITU-R Rec. BT.1122-2, "User requirements for codecs for emission and secondary distribution systems for SDTV and HDTV," March 2011.

Toshiko Tominaga received her B.E. and M.E. degrees in electrical engineering from University of Electro-Communications in Tokyo in 1987 and 1989. She joined NTT laboratories in 1989 and has been engaged in research into the quality assessment of facsimile and video. She is currently working on the quality assessment of video streaming and video-communication services over IP networks. She received the Telecommunication Advancement Foundation Award in Japan in 2008.

Masataka Masuda received his B.E. and M.E. degrees in electrical engineering from Shibaura Institute of Technology in Tokyo in 1997 and 1999, and Ph.D. degree in engineering from the Tokyo University of Agriculture and Technology in Japan in 2011. He joined NTT laboratories in 1999 and has been engaged in IP networks and the speech quality evaluation of VoIP. He is currently working on operations support architecture research for IP-based services. He received the Young Investigators' Award (IEICE) in Japan in 2002 and the Technical Committee on Communication Quality's Award (IEICE) in Japan in 2004.

Jun Okamoto received his B.E. and M.E. degrees in electrical engineering from Science University of Tokyo in Japan in 1994 and 1996. He joined NTT laboratories in 1996 and has been engaged in the quality assessment of visual communication services. Currently, he is working on subjective and objective assessment methods of audio and visual communication services. He has been contributing to VQEG (Video Quality Experts Group) since 2004. He received the Telecommunications Advancement Foundation Award in 2009 and the International Telecommunication Union Encouragement Award in Japan in 2010.

Akira Takahashi received his B.S. degree in mathematics from Hokkaido University in Japan in 1988, M.S. degree in electrical engineering from the California Institute of Technology in the U.S. in 1993, and Ph.D. degree in engineering from the University of Tsukuba in Japan in 2007. He joined NTT Laboratories in 1988 and has been engaged in the quality assessment of audio and visual communications. Currently, he is the Manager of the IP Service Network Engineering Group at NTT Laboratories. He was a co-Rapporteur of ITU-T Question 13/12 on Multimedia QoE and its assessment during the 2004–2008 Study Period. He is a Vice-Chairman of ITU-T Study Group 12 (SG12) for the 2009–2012 and 2013–2016 Study Periods. He received the Telecommunication Technology Committee Award in Japan in 2004 and the ITU-AJ Award in Japan in 2005. He also received the Best Tutorial Paper Award from IEICE in Japan in 2006 and the Telecommunications Advancement Foundation Award in Japan in 2007 and 2008.

Takanori Hayashi received his B.E., M.E., and Ph.D. degrees in Engineering from the University of Tsukuba, Ibaraki, in 1988, 1990, and 2007. He joined NTT Laboratories in 1990 and has been engaged in the quality assessment of multimedia telecommunication and network performance measurement methods. Currently, he is the Manager of the Service Assessment Group in NTT Laboratories. He received the Telecommunication Advancement Foundation Award in Japan in 2008 and the Telecommunication Technology Committee Award in Japan in 2012.