INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 1997 - 2000

Delayed Contribution **D-** _
January 1998
Original: English

Question: 11/12

SOURCE*: AT&T

TITLE: Examination of Agreement Between Laboratories Conducting Identical Subjective Video Quality Experiments

_____

ABSTRACT

The criteria to evaluate objective measures of video will be defined by the Video Quality Experts Group (VQEG). This contribution is an input to the discussion. It describes the lab-to-lab comparison of subjective test data collected as part of the T1A1.5 video performance research project. This project required a large set of video sequences with viewer ratings, and several laboratories shared the work. Low lab-to-lab MOS rms errors (<0.3) and high correlation (all r >0.94, most r >0.96) summarize these comparisons. The value of rms error as a criterion for comparisons is discussed.

* Contact persons - Alfred C. Morton Tel.: +1 732 949 2499, Fax -1652
E-Mail: acmorton@att.com

Introduction

A group of ITU Video Quality Experts met in Turin, Italy on October 14-16, 1997 to organize and plan the evaluation of objective video quality methods of measurement. Methods that show a strong relationship with viewer's opinion will be considered further in the development of ITU Recommendations. The criteria to evaluate these methods will soon be defined by the experts, and this is an input to the discussion.

This contribution describes the lab-to-lab comparison of subjective test data collected as part of the T1A1.5 video performance research project. This project required a large set of video sequences with viewer ratings. Several laboratories shared the work of collecting viewer opinions. The following analysis was conducted using raw data provided by the three test laboratories, GTE Labs, NTIA/ITS, and DIS/NCS.

The T1A1.5 Working Group defined 25 Hypothetical Reference Circuits (HRC) and 25 different test scenes, yielding 625 combinations. The plan called for 30 viewers to rate each HRC-scene combination in a double stimulus, impairment scale task (degradation category ratings). No subject could view all combinations, due to the extensive time requirements. Therefore, each viewer was assigned to one of three teams (designated Red, Green or Orange), and the combinations were divided evenly among the teams (with some overlap). Table 1 describes the HRCs and the division of HRCs among teams. The video test scenes were described in a contribution during the 1992-1996 study period (see ANSI T1.801.01-1994).

Each lab recruited at least 10 viewers for each team, permitting the ratings from any lab to be compared with the ratings from two identical experiments conducted at different locations. All lab facilities conformed with ITU-R Rec. 500 viewing conditions to great extent.

Data Preparation

We processed the raw data in accordance with the Subjective Test Plan (T1A1.5/94-118 R1), to obtain the following attributes:

1. The complete scores of viewers who failed to pass the vision acuity test, the color discrimination test, or the consistency checks, have been excluded. (except as approved by the Working Group).
2. For viewers who remain valid following these tests, all scores for the repeated trial and null HRC consistency checks have been excluded.
3. The Red, Green, and Orange Teams at each lab have no more than 10 viewers (any extra valid viewers were removed).

This process resulted in a data set containing the votes of 88 viewers.

Results of Lab to Lab Comparison

Table 2 Summary of Lab to Lab Comparisons

Fig.	Description	*r*	r^2	rms(diff)	rms(resid)
1	MOS GTE vs. ITS	0.966	0.933	0.302	0.286
2	GTE vs.ITS no Green	0.977	0.954	0.265	0.244
	MOS DIS vs. ITS	0.961	0.924	0.309	0.304
	DIS vs.ITS no Green	0.972	0.945	0.272	0.266
	MOS GTE vs. DIS	0.941	0.885	0.403	0.353
	GTE vs.DIS no Green	0.967	0.936	0.316	0.268

Figure 1 shows a comparison of Mean Opinion Scores (MOS) for each of the 625 test combinations, plotted as the GTE MOS (mosg) vs the ITS MOS (mosi). The correlation coefficient (r) is 0.966, indicating fairly good correlation between the two labs' results. The rms difference (or error) between mosg and mosi is 0.302.

The dashed lines illustrate a difference between the GTE and ITS data of 0.5 MOS. 53 of the 625 MOS differed more than 0.5 between labs.

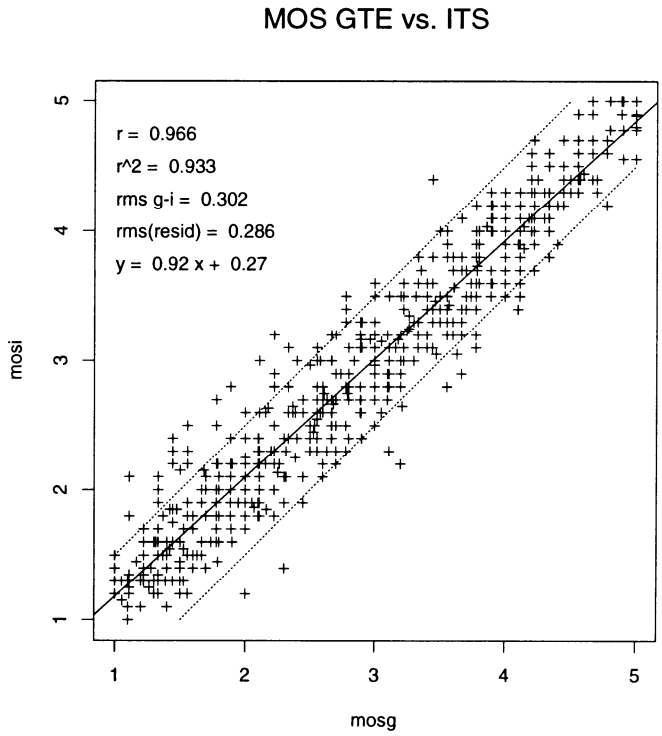

Figure 1

It appears that many of the large differences (>0.5 MOS) were combinations that ITS subjects rated as much as a full MOS point above the GTE subjects. The linear regression for this

comparison, taking mosi as the dependent variable yields y=0.92x+0.27, reflecting the effect of this bias.

Further investigation indicates that the limited bias occurred predominantly on HRCs viewed by the Green Team alone. Figure 2 shows the GTE and ITS MOS with all HRCs that the Green Team viewed removed (including HRCs viewed by more than one team). Coefficient r improves to 0.977, while the rms difference reduces to 0.265 MOS which also shows improvement. The regression line parameters also move closer to the ideal y=x.

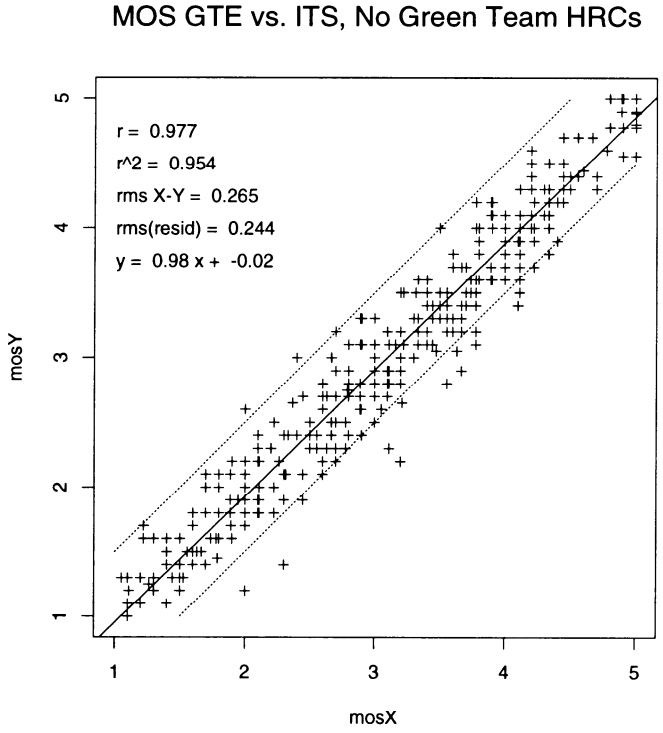

Figure 2

Table 2 shows similar lab to lab comparisons for the DIS vs. ITS and GTE vs. DIS. Again, we see improvement in correlation and rms difference measures with the Green Team HRCs excluded.

Discussion of Lab-Lab Comparisons

The results indicate that GTE, DIS/NCS and NTIA/ITS have consistently conducted the Subjective Test Plan, and delivered opinion data with high quality. There is at least one question remaining to be dealt with in the analysis: Why does the Green Team data contribute additional inconsistency? Potential factors may be found through examination of the viewer demographics or detailed examination of the HRCs assigned to this team.

We note that comparisons in terms of rms difference (or error) add value to the analysis. rms error retains the units of MOS, and permits interpretation of the results on a familiar scale where intervals have clear meaning.

The Lab to Lab Comparison results represent an important benchmark with which to compare the candidate objective estimates of MOS. Since the Lab MOS were derived from tests with 10 (or fewer) viewers, the benchmark is not an upper limit. Better consistency would be expected with larger viewer sample sizes, unless some systematic difference exists between labs. A sample of 10 viewers is a small experiment; it is less than the minimum of 15 given in ITU-R Rec. 500 and much less than sample sizes of 30 to 40 that are prevalent in subjective testing.

Conclusions

We conclude that the error distribution, whether expressed as raw score difference or regression residuals, is an important aid when assessing the closeness of comparisons. Examination of rms error permits interpretation of the results on a familiar scale.

The comparison indicates that Labs consistently followed the T1A1.5 Test Plan and delivered opinion data with high quality. Further investigation of the Green Team attributes would be informative.

We observe that the Lab to Lab Comparison results represent an important benchmark with which to compare the candidate objective measures of in terms of their correlation with Difference MOS**.** Since Lab to Lab correlation were derived from small experiments with 10 (or fewer) viewers, a benchmark based on their consistency would allow some uncertainty as an acceptance criteria for objective measures. Low lab-lab MOS rms errors (<0.3) and high correlation (all r >0.94, most r >0.96) summarize these comparisons.

TABLE 1 HYPOTHETICAL REFERENCE CIRCUITS and TEAM ASSIGNMENTS

HYPOTHETICAL REFERENCE CIRCUITS

These tables are a part of document TIAI.5/94-118 R1, Subjective Test Plan. The Testing Ad Hoc Group (H. Meiseles, Vyvx, Chair; S. Gallaher, Vyvx; A. Morton, AT&T Communications) prepared this table to describe the HRCs created by the Group using equipment available at the test site.

HRC	Algorithm (vendor)	Resolution	Total, Kbps	Audio, Kbps	Video, Kbps	Coding Mode	FEC	Burst Errors
1	Null			-	-		-	
2	VHS		-	-	-		-	
3	Proprietary	V.High	45,000	-	-		-	
4	Proprietary	Med.	128	-	-	VQ	-	
5	Proprietary	High	336	-	-	VQ	-	
6	Proprietary	Med.	112	-	-	-	-	
7	Proprietary	Med.	384	-	-	-	-	
8	Proprietary	Med.	768	-	-	-	-	
9	Proprietary	High	768	-	-	-	-	
10	Proprietary	High	1536	-	-	-	-	
11	H.261 (diff)	QCIF	128	56	70.4	INTER+MC	On	
12	H.261(same)	QCIF	128	56	70.4	INTER	On	
13	H.261(same)	QCIF	168	48	118.4	INTER+MC	On	
14	H.261(diff)	QCIF	384	56	326.4	INTER+MC	On	
15	H.261 (same)	CIF	112	48	62.4	INTER+MC	On	
16	H.261(same)	CIF	128	56	70.4	INTER+MC	On	
17	H.261(diff)	CIF	128	48	78.4	INTER+MC	On	
18	H.261(same)	CIF	168	48	118.4	INTER+MC	On	
19	H.261(same)	CIF	256	56	190.4	INTER+MC	On	On
20	H.261(same)	CIF	384	56	326.4	INTER+MC	On	
21	H.261(same)	CIF	384	56	326.4	INTER+MC	On	On
22	H.261(diff)	CIF	768	56	710.4	INTER+MC	On	
23	H.261(same)	CIF	768	56	710.4	INTER+MC	On	On
24	H.261(diff)	CIF	1536	56	1478.4	INTER+MC	On	
25	H.261(same)	CIF	1536	56	1478.4	INTER+MC	On	

TEAM - HRC ASSIGNMENTS

Red Tape Set:	1, 4, 7, 8,13,15,19,20,22,24
Green Tape Set:	2, 5, 6,10,14,15,16,17,20,23
Orange Tape Set:	3, 4, 9,11,12,17,18,20,21,25