JRG008

Contribution to Turin Experts Meeting on Subjective and Objective Audiovisual Quality

SOURCE[1]: National Telecommunications and Information Administration
Institute for Telecommunication Sciences
D. Atkinson, C. Jones

TITLE: Preliminary Results, Subjective Desktop Video Teleconferencing Audiovisual Test

Abstract

Preliminary results from a subjective desktop video teleconferencing audiovisual experiment performed at the Institute for Telecommunication Sciences (ITS) are discussed. This experiment was a single-ended audiovisual subjective test. Also presented are some technical considerations encountered in the setup and execution of this experiment.

---

1 Introduction

This experiment was conducted to collect audiovisual subjective data for representative audiovisual desktop video teleconferencing systems. It was designed such that audio-only and video-only subjective data would also be collected. With these three sets of subjective data (audiovisual, audio only, and video only), we will be able to map audio-only and video-only subjective scores into an overall audiovisual subjective score.

Staff at the Institute for Telecommunication Sciences (ITS) will also be developing objective models to predict the audio-only subjective scores and the video-only subjective scores. This will be done with measurements and methods previously developed at ITS. In addition, ITS staff will develop a preliminary objective model of audiovisual subjective scores using the results of the individual objective audio and video models.

This paper discusses the design of the audiovisual experiment, the test procedure used, and gives some preliminary subjective data. Also discussed are some topics that the Study Groups concerned with developing a subjective audiovisual testing recommendation might take into consideration.

---

[1] Contact: Coleen Jones, NTIA/ITS, +1-303-497-3764,
Email: cjones@its.bldrdoc.gov

2 Test Plan

The primary goal of this test was collection of subjective performance data for representative desktop video teleconferencing (DVTC) applications. As a secondary goal, this data will support the development of an objective audiovisual quality model. The audiovisual quality model will be some combination of the individual objective audio and video measurements developed at ITS. A model of audiovisual subjective quality based upon the individual audio and video subjective scores will also be investigated.

This test included typical DVTC equipment such as a computer monitor and desktop computer speakers, but it took place in an acoustically isolated chamber. The audio and video was processed through several representative DVTC configurations.

This test consisted of three individual sessions, a video only test, an audio only test, and a combined audiovisual test. The order of presentation of these three tests was permuted for each test subject.

2.1 Test Design

Table 1 lists the design parameters in this test. A source tape in the component BetacamSP format was used as input to each of the eight processing configurations listed in Table 2. Both the input and output of the configurations was composite (NTSC) video, since this is a likely format to be used by DVTC users. Because we wanted to remove delay as a factor in the audiovisual quality rating, the audio was delayed such that the audio and video were synchronized. The delay was fixed for each condition, and it is listed in Table 2. The NTSC output of the configurations was recorded in BetacamSP format and played back to the subjects in S-video (component Y/C) format. The video was input to a PC overlay card and displayed on a PC monitor for the subjects to view. The audio was delivered via typical PC speakers. The performance ratings were gathered using the absolute category rating (ACR) method for all three sessions [1].

The six scenes were selected as representative examples of typical VTC scenes. The scenes vtc1nw and smity2 consist of one person (vtc1nw has very little motion, and smity2 has a moderate amount of motion). The scene vtc2 has one person with graphics (a map). The first portion of this scene has little motion, and the second portion of this scene has a camera zoom that creates a lot of motion. The scene 5row1 has five people sitting around a conference table. And filter and washdc are two graphics-related scenes.

Each of the six scenes was run through each of the eight processing configurations. Each of the 18 subjects was presented all 48 conditions in each session. However, each subject received one of six rating session permutations (e.g. video only first, audio only second, and audio and video third), resulting in each of the six permutations being rated by 3 subjects.

Table 1 Test Design Parameters

Design Parameter	
Type of test:	ACR, in 3 sessions, video only, audio only, and combined audio/video
Scenes:	5row1 (2 talkers), filter, smity2, vtc1nw, vtc2, washdc (2 talkers)
Viewing device:	17" PC Monitor
Viewing distance:	Approximately 4-8 times the video window height
Listening device:	PC Multimedia speakers
Subjects:	18, chosen from US Dept of Commerce, Boulder Labs staff.

Table 2 Test Processing Configurations

Condition Number	System	Video Algorithm	Audio Algorithm	Delay (ms)
1	NTSC (525-line Composite)	Analog (NTSC)	Analog	0
2	1536 kb/s, System A	H.261 CIF (1472 kb/s)	G.722 64	80
3	1536 kb/s, System B	Prop. Alg. A (1472 kb/s)	G.722 64	16
4	384 kb/s, System B	H.261 QCIF (320 kb/s)	G.711 64	100
5	384 kb/s, System A	H.261 CIF (320 kb/s)	G.722 64	120
6	128 kb/s, System A	H.261 QCIF (112 kb/s)	G.728 16	200
7	128 kb/s, System B	H.261 QCIF (64 kb/s)	G.711 64	144
8	128 kb/s, System B	Prop. Alg. B (120 kb/s)	Prop. Alg. (8 kb/s)	30

3 Preliminary Results

The per-clip mean opinion score (MOS), averaged over 18 subjects, is listed for each test session (audio-only, video-only, and audiovisual) in Table 4, Table 5, and Table 6 respectively. The mean opinion scores are also plotted in Figure 1.

When looking at the audio mean opinion scores for scenes (averaged over subjects and conditions, i.e. the scene main effect), there is more variation between scenes within this data set than is typically seen in an ACR subjective audio quality test. This can mainly be attributed to variable levels of audio quality in the source material. Two scenes had good quality audio tracks (filter and washdc), and the other four scenes (5row1, smity2, vtc1nw, vtc2) had lower quality, noisy audio tracks. This variation in source quality is difficult to account for using an ACR test. Perhaps a degradation category rating (DCR) test would have resulted in less variable audio quality scores for this test. However, the testing parameters (mainly test length) constrained us to running an ACR test.

The confidence intervals on the video and audiovisual mean opinion scores are reasonable. However, the confidence intervals on the audio mean opinion scores are larger than typically found in audio ACR tests. This is most likely due to the variation in source audio quality as discussed above.

For the 48 clip mean opinion scores shown in Tables 4-6, the correlation coefficients between the different tests are listed in Table 3 below.

Table 3 Between-test correlation coefficients

$\rho_{a,v}$: 0.29

$\rho_{a,av}$: 0.41

$\rho_{v,av}$: 0.97

It appears that, for the case of the video teleconferencing systems in this test, video quality seems to be the main factor of audiovisual quality.

Upon initial investigation, there appears to be no significant ordering effects due to the rating session permutations of audio only, video only, and audiovisual. A more detailed analysis of the ordering effects will be undertaken in the future to verify the significance of any session ordering effects.

It is interesting to note the difference between the video mean opinion scores for the scene vtc1nw for the first three conditions (NTSC, and two 1536 kb/s systems, see clips 4, 10, and 16 in Table 5). One would expect that the NTSC video would receive a higher MOS than the two 1536 kb/s-coded video scenes. What we are seeing here is the effect of the overlay card used to display the video on a PC monitor. The overlay card digitized the video with 8-bit color before displaying it on the PC monitor. The NTSC video (for this specific scene, vtc1nw) exhibited poor color quantization effects. There were areas on the woman's face that were much lighter than her surrounding skin tone. This problem did not occur with the two 1536 kb/s-coded video scenes, causing them to be rated higher than the NTSC video scene. Thus, for this scene, the overlay card affected the video quality ratings more than the coding methods.

Table 4 Audio Session Subjective Scores

Clip Number	Condition Number	Scene Name	Audio Session		
			MOS	Sample Std. Deviation	95% Half-width Confidence Interval
1	1	5row1	3.389	0.979	0.452
2	1	filter	4.611	0.608	0.281
3	1	smity2	3.667	0.970	0.448
4	1	vtc1nw	2.889	1.023	0.472
5	1	vtc2	3.278	1.074	0.496
6	1	washdc	4.444	0.705	0.326
7	2	5row1	3.167	0.786	0.363
8	2	filter	4.611	0.502	0.232
9	2	smity2	3.556	0.784	0.362
10	2	vtc1nw	2.667	0.907	0.419
11	2	vtc2	3.278	1.018	0.470
12	2	washdc	4.056	0.639	0.295
13	3	5row1	3.444	0.922	0.426
14	3	filter	3.500	0.786	0.363
15	3	smity2	3.389	1.037	0.479
16	3	vtc1nw	2.611	0.778	0.359

Clip Number	Condition Number	Scene Name	Audio Session		
			MOS	Sample Std. Deviation	95% Half-width Confidence Interval
17	3	vtc2	3.000	0.767	0.354
18	3	washdc	3.278	1.018	0.470
19	4	5row1	3.222	0.943	0.436
20	4	filter	2.889	0.758	0.350
21	4	smity2	3.389	0.850	0.393
22	4	vtc1nw	2.611	0.778	0.359
23	4	vtc2	3.056	0.725	0.335
24	4	washdc	3.389	0.698	0.322
25	5	5row1	3.389	1.037	0.479
26	5	filter	4.611	0.608	0.281
27	5	smity2	3.778	1.060	0.490
28	5	vtc1nw	2.722	0.669	0.309
29	5	vtc2	3.278	0.669	0.309
30	5	washdc	4.167	0.924	0.427
31	6	5row1	2.833	0.857	0.396
32	6	filter	4.222	0.808	0.373
33	6	smity2	3.167	0.985	0.455
34	6	vtc1nw	2.722	0.669	0.309
35	6	vtc2	2.500	0.924	0.427
36	6	washdc	3.444	0.784	0.362
37	7	5row1	3.056	0.938	0.433
38	7	filter	3.500	0.707	0.327
39	7	smity2	3.556	1.042	0.481
40	7	vtc1nw	2.833	0.707	0.327
41	7	vtc2	3.056	0.725	0.335
42	7	washdc	3.333	0.840	0.388
43	8	5row1	1.889	0.758	0.350
44	8	filter	2.889	0.583	0.269
45	8	smity2	1.833	0.707	0.327
46	8	vtc1nw	1.778	0.548	0.253
47	8	vtc2	1.444	0.511	0.236
48	8	washdc	2.389	0.698	0.322

Table 5 Video Session Subjective Scores

Clip Number	Condition Number	Scene Name	Video Session		
			MOS	Sample Std. Deviation	95% Half-width Confidence Interval
1	1	5row1	4.389	0.608	0.281
2	1	filter	4.611	0.608	0.281
3	1	smity2	4.556	0.511	0.236
4	1	vtc1nw	3.889	1.132	0.523
5	1	vtc2	4.667	0.485	0.224
6	1	washdc	4.667	0.485	0.224
7	2	5row1	3.722	0.752	0.347
8	2	filter	3.889	0.676	0.312
9	2	smity2	3.667	0.767	0.354
10	2	vtc1nw	4.333	0.594	0.274
11	2	vtc2	3.556	0.922	0.426
12	2	washdc	3.722	0.575	0.265
13	3	5row1	4.222	0.548	0.253
14	3	filter	4.056	0.539	0.249
15	3	smity2	3.333	0.767	0.354
16	3	vtc1nw	4.222	0.732	0.338
17	3	vtc2	3.389	0.608	0.281
18	3	washdc	3.833	0.618	0.286
19	4	5row1	1.667	0.594	0.274
20	4	filter	2.722	1.274	0.589
21	4	smity2	2.444	0.784	0.362
22	4	vtc1nw	2.389	0.502	0.232
23	4	vtc2	2.111	0.676	0.312
24	4	washdc	1.611	0.502	0.232
25	5	5row1	3.444	0.705	0.326
26	5	filter	3.667	0.594	0.274
27	5	smity2	2.889	0.832	0.385
28	5	vtc1nw	3.889	0.583	0.269
29	5	vtc2	2.611	0.979	0.452
30	5	washdc	3.556	0.705	0.326
31	6	5row1	1.333	0.485	0.224
32	6	filter	1.833	0.618	0.286
33	6	smity2	1.889	0.676	0.312
34	6	vtc1nw	2.056	0.802	0.371
35	6	vtc2	1.000	0.000	0.000
36	6	washdc	1.167	0.383	0.177

Clip Number	Condition Number	Scene Name	Video Session		
			MOS	Sample Std. Deviation	95% Half-width Confidence Interval
37	7	5row1	1.611	0.608	0.281
38	7	filter	1.722	0.669	0.309
39	7	smity2	1.278	0.461	0.213
40	7	vtc1nw	2.167	0.707	0.327
41	7	vtc2	1.000	0.000	0.000
42	7	washdc	1.333	0.485	0.224
43	8	5row1	2.778	0.647	0.299
44	8	filter	3.333	0.594	0.274
45	8	smity2	1.722	0.575	0.265
46	8	vtc1nw	3.611	0.608	0.281
47	8	vtc2	2.167	0.707	0.327
48	8	washdc	2.778	0.732	0.338

Table 6 Audiovisual Session Subjective Scores

Clip Number	Condition Number	Scene Name	Audiovisual Session		
			MOS	Sample Std. Deviation	95% Half-width Confidence Interval
1	1	5row1	4.222	0.732	0.338
2	1	filter	4.667	0.485	0.224
3	1	smity2	4.722	0.461	0.213
4	1	vtc1nw	3.778	1.215	0.561
5	1	vtc2	4.389	0.698	0.322
6	1	washdc	4.556	0.705	0.326
7	2	5row1	3.722	0.752	0.347
8	2	filter	4.278	0.752	0.347
9	2	smity2	3.944	0.873	0.403
10	2	vtc1nw	3.833	0.707	0.327
11	2	vtc2	3.722	0.826	0.382
12	2	washdc	3.667	0.686	0.317
13	3	5row1	4.278	0.752	0.347
14	3	filter	4.111	0.758	0.350
15	3	smity2	3.500	0.924	0.427
16	3	vtc1nw	3.722	0.669	0.309
17	3	vtc2	3.444	0.705	0.326
18	3	washdc	3.778	0.647	0.299

Clip Number	Condition Number	Scene Name	Audiovisual Session		
			MOS	Sample Std. Deviation	95% Half-width Confidence Interval
19	4	5row1	1.889	0.758	0.350
20	4	filter	2.778	1.114	0.515
21	4	smity2	2.889	0.676	0.312
22	4	vtc1nw	2.167	0.786	0.363
23	4	vtc2	2.056	0.725	0.335
24	4	washdc	1.889	0.583	0.269
25	5	5row1	3.389	0.916	0.423
26	5	filter	3.944	0.539	0.249
27	5	smity2	2.111	1.023	0.472
28	5	vtc1nw	3.667	0.767	0.354
29	5	vtc2	2.333	1.085	0.501
30	5	washdc	3.722	0.826	0.382
31	6	5row1	1.500	0.618	0.286
32	6	filter	1.833	0.707	0.327
33	6	smity2	1.611	0.608	0.281
34	6	vtc1nw	2.111	0.583	0.269
35	6	vtc2	1.111	0.323	0.149
36	6	washdc	1.222	0.428	0.198
37	7	5row1	1.778	0.647	0.299
38	7	filter	1.778	0.808	0.373
39	7	smity2	1.556	0.616	0.284
40	7	vtc1nw	2.500	0.707	0.327
41	7	vtc2	1.167	0.383	0.177
42	7	washdc	1.722	0.752	0.347
43	8	5row1	2.389	0.850	0.393
44	8	filter	3.111	0.758	0.350
45	8	smity2	1.611	0.502	0.232
46	8	vtc1nw	2.778	1.003	0.463
47	8	vtc2	2.056	0.802	0.371
48	8	washdc	2.444	0.856	0.395

4 Subjective Audiovisual Testing Considerations

Several testing details were encountered during preparation for this test. They are listed and explained below as topics for discussion concerning the development of an audiovisual subjective testing recommendation.

- Audio Availability - When we were selecting source material, we noticed that the ANSI standardized test scenes had either no audio or low quality audio. We attempted to compensate for this by selecting additional test scenes that had better audio, and, in one case, recording a new audio track. We would like to urge the experts group to consider the quality of the audio along with the quality of the video in the source material when selecting test scenes for Recommendation.
- Lighting – Rec. 500-5 lists the ratio of luminance of background behind picture monitor to peak luminance of picture to be approximately 0.15 with a chromaticity of D_{65}. Rec. P.910 lists the ratio of background luminance to maximum screen luminance to be approximately 0.25 with no chromaticity listed. What is the rationale behind these numbers and their differences? Does one make more sense than the other? If a window within the screen is used, is this ratio still significant? Chromaticity of D_{65} is difficult to obtain without using studio-quality lighting.
- Display Device – Display device should be specified, or perhaps an option of several display devices given. For example, interactive tests might be tested using PC monitors as in this test. If PC monitors are used, several additional considerations arise. Should video be displayed within a window? If so, what is an appropriate background luminance level? Also, is the ratio between background color and peak luminance within the window significant? Is it more significant than the room-to-monitor luminance ratio? What VGA resolution is recommended? What size of monitor is appropriate? At what color temperature should the PC monitor be set (assuming it is adjustable)? Our PC monitor could be adjusted to a color temperature of 9300, 6500, or 5000. We chose 6500, consistent with Rec. 500-5. What type of display device is suggested for display on a PC monitor? NTSC to VGA scan converter? Video overlay card? We used an 8-bit color depth overlay card. The overlay card itself significantly affected some of our video scenes (see Section 3).
- Speakers – What type of listening device is recommended? A handset, headphones, PC speakers, loudspeakers? We chose PC speakers to be consistent with typical DVTC uses.
- Environment – What type of environment is appropriate? An acoustically isolated room? A "typical" office environment? A "typical" lab environment? What noise characteristics should be used to simulate these environments?
- Background noise – When might background noise be useful and what type and level is recommended?
- Video format – What video formats are suggested, composite (NTSC, PAL), component s-video (Y/C), component (RGB, BetacamSP)?

5 Summary

This paper offers preliminary results of an audiovisual subjective test that contained an audio-only session, a video-only session, and an audiovisual session. The subjective data has revealed that careful attention should be paid to the quality of the source audio and the effects of display devices such as overlay cards (or any piece of equipment that digitizes or converts video). The subjective data has also revealed that for video teleconferencing systems (rates at or below 1536 kb/s), the quality of the video is the main factor in the audiovisual quality rating. It also appears that the ordering of the rating sessions (audio only, video only, audiovisual) does not affect the mean opinion scores.

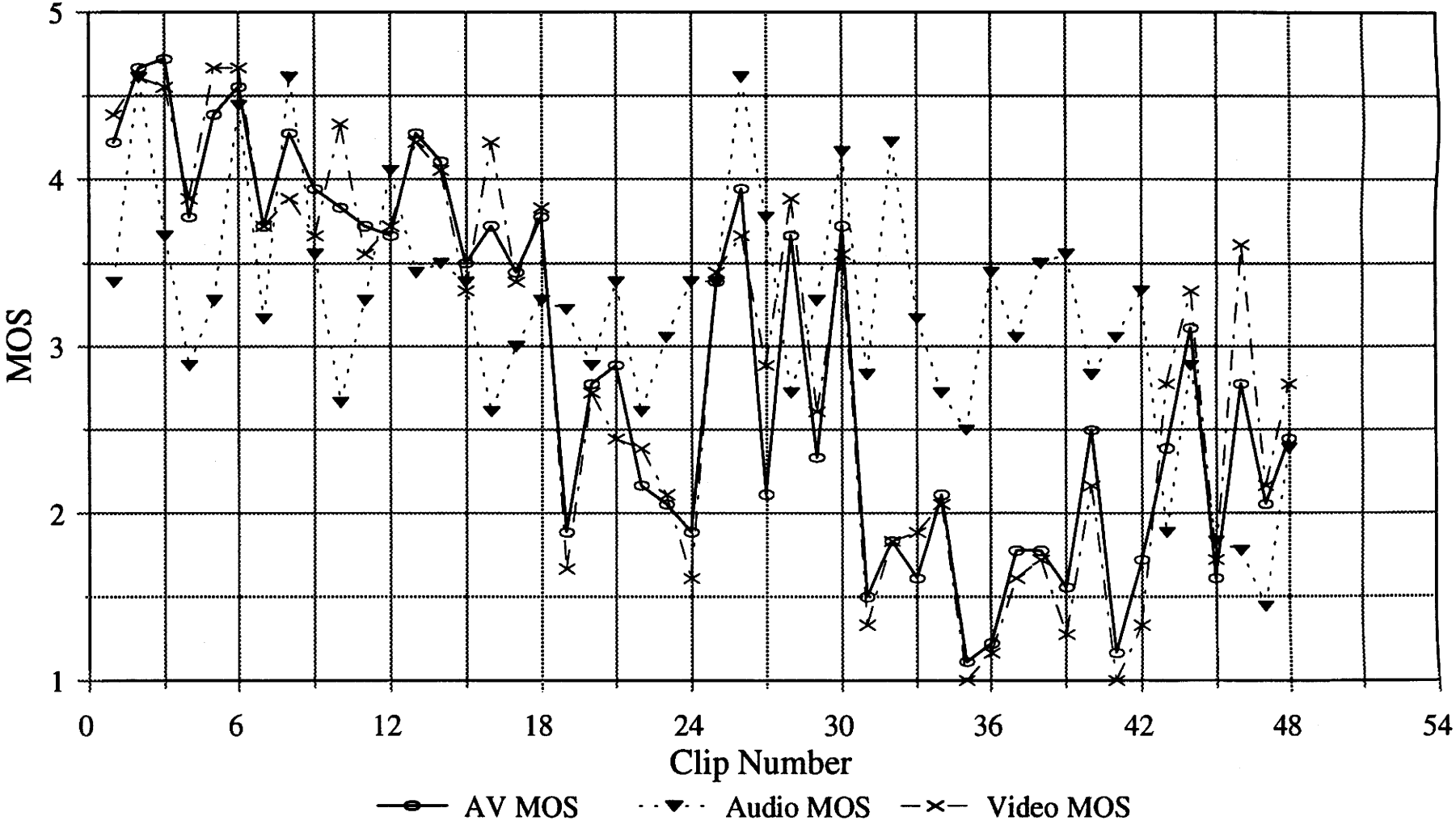

Figure 1. Clip mean opinion score for audio-only test, video-only test and audiovisual test.
(see Table 6 to relate clip number to testing condition and scene)

6 References

[1] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications", Recommendations of the ITU (Telecommunication Standardization Sector).