UIT - Secteur de la normalisation des télécommunications
ITU - Telecommunication Standardization Sector
UIT - Sector de Normalización de las Telecomunicaciones

Commission d'études ;Study Group;Comisión de Estudio} 12
Contribution tardive;Delayed Contribution ;Contribución tardía} **D.xxx**

Texte disponible seulement en ;Text available only in;Texto disponible solamente en} E

Question: 10/12

SOURCE: USA[1]

TITLE: Results of an audiovisual desktop video teleconferencing subjective experiment

Abstract

This contribution discusses the analysis of an audiovisual desktop video teleconferencing subjective experiment conducted at the Institute for Telecommunication Sciences in Boulder, Colorado, USA. Results of the subjective data analysis, including session ordering effects are presented. Also, a subjective model of audiovisual quality based upon the individual subjective audio and video scores is discussed.

[1] Contact: Ms. Coleen Jones, Tel: +1 303 497 3764 Fax: +1 303 497 5323
E-mail: cjones@its.bldrdoc.gov

1. Introduction

The Institute for Telecommunication Sciences (ITS) conducted an audiovisual desktop video teleconferencing subjective experiment with the goal of investigating the relationship between the individual audio and video quality and the overall audiovisual quality. The experiment is explained, and some suggestions are made that Study Group 12 might take into consideration while developing Recommendations in this area.

Using the subjective data, we have developed a subjective audiovisual quality model that relates the individual subjective audio and video data to the overall subjective audiovisual quality. We have also compared our results with subjective audiovisual models developed by other laboratories.

A description of the subjective audiovisual experiment and the results of the subjective data analysis are presented.

This contribution describes work in progress. As a summary of an initial study, it is offered for consideration during the preparation of a draft new recommendation on combined audiovisual quality subjective assessment. Further study of this topic is needed.

2. Test Plan

The primary goal of this test was the collection of subjective performance data for representative desktop video teleconferencing (DVTC) applications. 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 were processed through several representative DVTC configurations.

This test consisted of three individual sessions:

- a video-only session in which subjects saw only video and rated the video quality;
- an audio-only session in which subjects heard only audio and rated the audio quality;
- an audiovisual session in which subjects rated the overall quality of an audiovisual clip.

A source tape in professional ½" component analog video format was used as input to each of the eight processing configurations listed in Table 1. Both the input and output of the configurations were composite (NTSC) video, because this reflects the typical 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 synchronised. The adjusted audiovisual delay for each configuration is listed in Table 1. The NTSC output of the configurations was recorded in professional ½" component video 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 17" PC monitor for the subjects to view. The audio was delivered via typical PC multimedia speakers. The performance ratings were gathered using the five-point absolute category rating (ACR) method for all three sessions [1].

The six test scenes selected were representative of VTC scenes [2]. 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. Filter and washdc are graphics-related scenes.

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

Table 1: Processing Configurations

Configuration Number	Processing Configuration	Video Algorithm[1]	Audio Algorithm[1]	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[2]	80
3	1536 kb/s, System B	Prop. Alg. A (1472 kb/s)	G.722	16
4	384 kb/s, System B	H.261 QCIF (320 kb/s)	G.711[3]	100
5	384 kb/s, System A	H.261 CIF (320 kb/s)	G.722	120
6	128 kb/s, System A	H.261 QCIF (112 kb/s)	G.728[4]	200
7	128 kb/s, System B	H.261 QCIF (64 kb/s)	G.711	144
8	128 kb/s, System B	Proprietary Algorithm B (120 kb/s)	Prop. Alg. (8 kb/s)	30

[1] Bit rates listed do not account for bits used for overhead/control information.
[2] G.722: 7 KHz bandwidth at 64 kb/s.
[3] G.711: 4 KHz bandwidth at 64 kb/s.
[4] G.728: 4 KHz bandwidth at 16 kb/s.

3. 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 audiovisual subjective testing Recommendations.

- We suggest that SG-12 consider the quality of the audio along with the quality of the video in the source material when selecting test scenes for Recommendation. High quality audio is as important as high quality video for achieving good results.
- There is some inconsistency between the lighting recommendations in ITU-R BT.500-5 and Rec. P.910. ITU-R BT.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. Careful thought should be given to the lighting requirements in the subjective audiovisual testing Recommendations being developed by SG-12. Also, thought should be given to screen background illumination if the video is presented within a window (of a PC application) on a PC monitor.
- The video display device is an important piece of equipment when video quality is being tested. It is suggested that the display device be specified, or perhaps an option of several display devices given. For example, interactive tests might be tested using PC monitors (as were used in this test). If PC monitors are used, several additional considerations arise. If the video is displayed within a window, what is an appropriate background luminance level? Also, is the ratio between background colour and peak luminance within the window significant? Is the ratio between background colour and peak luminance more significant than the room-to-monitor luminance ratio?

  Other considerations are VGA resolution, monitor size, and the colour temperature of the PC monitor. Our PC monitor could be adjusted to a colour temperature of 9300, 6500, or 5000. We chose 6500, consistent with ITU-R BT.500-5. If the video is displayed on a PC monitor, some type of conversion device will be necessary. Some possibilities are a composite (PAL or NTSC) to VGA scan converter or a video overlay card. We used an 8-bit colour depth overlay card. The overlay card itself significantly affected some of our video scenes (see Section 4).
- Several options are available for delivery of the audio, such as a handset, headphones, PC speakers, or loudspeakers. We chose PC speakers to be consistent with typical DVTC uses.

- It is suggested that SG-12 consider different testing environments. Several environments could be covered. For example, acoustically isolated rooms, a "typical" office environment, or a "typical" lab environment.
- It would be interesting to define several background noise sources that could be published on CD-ROM. For example, "office noise" at several levels or "lab noise" at several levels. A CD-ROM of this type could be useful for both subjective audiovisual testing and audio testing.
- Video format will most likely be laboratory or experiment-specific. However, some general guidelines would be useful, especially for the display of the video on a monitor or other video device such as a PC monitor.

4. Subjective Results

Averaging over all subjects (18) for each scene-processing configuration combination yields what we term a clip MOS. Figure 1 is a plot of the clip MOS for all 48 conditions (6 scenes and 8 processing configurations). The clip MOS is plotted for the audio-only session (▼), the video-only session (×), and the audiovisual session (o). Table 2 relates clip number to processing configuration and scene name.

Table 2: Relationship between clip number, configuration number, and scene name.

Clip Number	Config. Number	Scene Name	Clip Number	Config. Number	Scene Name	Clip Number	Config. Number	Scene Name
1	1	5row1	17	3	vtc2	33	6	smity2
2	1	filter	18	3	washdc	34	6	vtc1nw
3	1	smity2	19	4	5row1	35	6	vtc2
4	1	vtc1nw	20	4	filter	36	6	washdc
5	1	vtc2	21	4	smity2	37	7	5row1
6	1	washdc	22	4	vtc1nw	38	7	filter
7	2	5row1	23	4	vtc2	39	7	smity2
8	2	filter	24	4	washdc	40	7	vtc1nw
9	2	smity2	25	5	5row1	41	7	vtc2
10	2	vtc1nw	26	5	filter	42	7	washdc
11	2	vtc2	27	5	smity2	43	8	5row1
12	2	washdc	28	5	vtc1nw	44	8	filter
13	3	5row1	29	5	vtc2	45	8	smity2
14	3	filter	30	5	washdc	46	8	vtc1nw
15	3	smity2	31	6	5row1	47	8	vtc2
16	3	vtc1nw	32	6	filter	48	8	washdc

It is interesting to note the difference between the video mean opinion scores for the scene vtc1nw for the first three configurations (NTSC (3.89), and two 1536 kb/s configurations (4.33 and 4.22), see clips 4, 10, and 16 in Figure 1). One would expect that the NTSC video would receive a higher MOS than the two 1536 kb/s-coded video scenes. This is an effect of the overlay card used to display the video on a PC monitor. The overlay card uses an 8-bit colour palette to display video on the PC monitor. In the NTSC video, the woman's cheeks were shiny, but due to processing in the two 1536 kb/s configurations, her cheeks appeared a normal skin tone. Thus, when the NTSC video was fed through the overlay card for display, it exhibited poor colour quantization effects resulting in unnatural skin tones. 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.

For the first six clips (NTSC processing configuration), the audio MOS varies by more than one and a half quality units (see Figure 1), which is larger than would normally be expected. The other processing configurations exhibit this pattern as well. The data corroborates that

two scenes had high quality audio tracks (filter and washdc), and the other four scenes (5row1, smity2, vtc1nw, vtc2) had lower quality audio (with background noise).

Table 3 lists the summary statistics of the confidence intervals for the three rating sessions. The confidence intervals on the video and audiovisual mean opinion scores are reasonable, being near 0.3 on average. However, the confidence intervals on the audio mean opinion scores are larger (nearly 0.4 quality units) than typically found in audio ACR tests. This is most likely due to the variation in source audio quality as discussed above.

Table 3: Confidence interval summary statistics for each of the rating sessions.

Session:	Min	Max	Mean	Median
Audio	0.232	0.496	0.373	0.362
Video	0.0	0.589	0.293	0.281
Audiovisual	0.149	0.561	0.338	0.337

For the 48 clip mean opinion scores shown in Figure 1, the correlation coefficients between the different sessions are listed in Table 4 below.

Table 4: Between-test correlation coefficients

Correlation between audio and video sessions($\rho_{a,v}$)	0.29
Correlation between audio and audiovisual sessions($\rho_{a,av}$)	0.41
Correlation between video and audiovisual sessions ($\rho_{v,av}$)	0.97

It appears that, for the case of the video teleconferencing systems in this test, video quality seems to be the main factor in audiovisual quality. These results are similar to those obtained by KPN [3] in a similar experiment that resulted in correlation coefficients of $\rho_{a,v} = -0.02$, $\rho_{a,av} = 0.33$, and $\rho_{v,av} = 0.90$.

We conducted an analysis to determine whether or not the session ordering was significant. We calculated the session MOS by averaging over all 48 conditions and all subjects who saw a given session either first, second, or third during their testing. For example, six subjects rated video in the third session (audio, audiovisual, video or audiovisual, audio, video). Thus, the session MOS for subjects that rated video third ($\bar{v}_3$) has been averaged over these 6 subjects and all 48 test conditions. We then calculated the session MOS differences (three differences each for the video sessions, audio sessions, and audiovisual sessions) to compare differences between rating video, audio, or audio-video first, second, or third. When the confidence interval for a difference does not span zero, that difference is deemed significant. Table 5 lists the session MOS differences and confidence intervals. The confidence intervals assume an approximate Gaussian distribution given the large number of samples over which we are averaging. The session MOS differences are also plotted in Figure 2.

Table 5 shows that the three audio session MOS differences are close to zero, indicating that for the audio sessions, there are no significant ordering effects. However, presentation order was significant in the other sessions. Subjects rated video and audiovisual quality higher (by about 0.2 to 0.3 quality units) in the third session compared to earlier sessions. The video session MOS differences are near zero at the bounds of the confidence interval. Thus, the video session MOS differences may be called marginally significant effects, but it remains that the differences are not as small as one would hope. The audiovisual session differences are more significant, even when the confidence intervals are taken into account. This may be due to subjects becoming accustomed to, and more tolerant of, the degraded video quality. Additional experimentation is necessary to determine the exact cause of these ordering effects and the experimental procedure that will minimize them.

Table 5: Session MOS Differences

	Session MOS Difference	Half-width Confidence Interval (95%)	Confidence Interval Bounds	
$\overline{v}_1 - \overline{v}_2$	-0.051	0.217	-0.268	0.166
$\overline{v}_1 - \overline{v}_3$	-0.222	0.218	-0.44	-0.004
$\overline{v}_2 - \overline{v}_3$	-0.172	0.202	-0.374	0.03
$\overline{a}_1 - \overline{a}_2$	-0.021	0.173	-0.194	0.152
$\overline{a}_1 - \overline{a}_3$	0.031	0.170	-0.139	0.201
$\overline{a}_2 - \overline{a}_3$	0.052	0.184	-0.132	0.236
$\overline{av}_1 - \overline{av}_2$	0.073	0.210	-0.137	0.283
$\overline{av}_1 - \overline{av}_3$	-0.257	0.213	-0.47	-0.044
$\overline{av}_2 - \overline{av}_3$	-0.330	0.211	-0.541	-0.119

5. Subjective Audiovisual Models

A model that relates the individual subjective audio and video mean opinion scores (s_a, s_v) to the subjective audiovisual mean opinion scores was investigated. Several different forms of equations were analyzed including cross products (between the audio and video terms), and the sums and differences of first and second order terms in different permutations. Two models had similar correlation coefficients when compared with the audiovisual subjective data as seen in Table 6, ITS models 1 and 3. The additional cross term in model 3 does not significantly improve the correlation coefficient; therefore we used the simpler linear combination in model 1. The model is shown in equation (1).

$$\hat{s}_{av} = -0.677 + 0.888 s_v + 0.217 s_a \quad (1)$$

The correlation coefficient (ρ) between the subjective audiovisual scores and the model in equation (1) is 0.978 ($\rho^2 = 0.957$). The scatter plot of the subjective audiovisual clip MOS versus the subjective audiovisual model output is shown in Figure 3. For comparison, the ITS model 3 output versus the subjective audiovisual clip MOS is shown in Figure 5.

Table 6: Comparison of models from different laboratories

Laboratory	Model	ρ	ρ^2
ITS	1: $\hat{s}_{av} = -0.677 + 0.217 s_a + 0.888 s_v$	0.978	0.957
	2: $\hat{s}_{av} = 1.514 + 0.121 (s_v \times s_a)$	0.927	0.859
	3: $\hat{s}_{av} = 0.517 - 0.0058 s_a + 0.654 s_v + 0.042 (s_a \times s_v)$	0.980	0.960
KPN	1: $\hat{s}_{av} = 1.45 + 0.11 (s_v \times s_a)$	0.97	0.94
	2: $\hat{s}_{av} = 1.12 + 0.007 s_a + 0.24 s_v + 0.088 (s_a \times s_v)$	0.98	0.96
Bellcore	1: $\hat{s}_{av} = 1.07 + 0.111 (s_v \times s_a)$	0.99	0.98
	2: $\hat{s}_{av} = 1.295 + 0.107 (s_v \times s_a)$	0.99	0.98

For the subjective model, all of the data was converted to a 9-point scale so that the models could be compared to models from other laboratories that conducted their subjective experiments using a 9-point ACR scale as opposed to the 5-point ACR scale used in this experiment. We typically consider prediction errors greater than one-half quality unit to be significant. However, because we have converted this data to a 9-point scale, prediction errors greater than one quality unit are deemed significant. All of the prediction errors for ITS model 1 are less than one quality unit, with the exception of clip 27 that exhibits a relatively large error of –1.77. Because we did not want audiovisual delay to be a factor in this experiment, we delayed the audio so that it would be synchronized with the video. We chose a fixed delay for each configuration (this delay is denoted in Table 1). However, for this specific combination of scene and processing configuration, the chosen delay was not accurate in conjunction with a significant amount of frame repetition. Thus, the audio and video were not synchronized, and subjects rated the audiovisual sequence worse than they rated the individual audio and video sequences. The subjective audiovisual model could not account for this difference.

Other laboratories have conducted similar experiments and developed subjective audiovisual models [3,4]. Table 6 summarizes results from ITS, KPN, and Bellcore. All three laboratories have investigated a model based upon the product of the individual audio and video subjective scores. All three laboratories have achieved similar results, with most of the variation seen in the additive constant. ITS model 2 did not correlate with the subjective audiovisual data as well as the KPN model 1 and Bellcore models 1 and 2. This may be due to either our noisy source material, or the different impairments used in our experiment. The subjective audiovisual clip MOS versus the ITS model 2 output is shown in Figure 4.

KPN model 2 adds the product term to the linear model. They found the interaction between audio and video to be significant (using an analysis of variance), and thus they included the product in their model. The constants in ITS model 3 and KPN model 2 are quite different, yet both models achieve the same correlation with the subjective audiovisual data. Note that for both models the audio quality factor is near zero. This is consistent with the low correlation coefficients between audio MOS and audiovisual MOS reported by both laboratories. The ITS audio factor is even negative which is counter-intuitive, and should be set to zero.

With ITS' objective models, we have found that the coefficients are dependent upon both the application and the population from which the subjective results were obtained. For example, broadcasters are much more critical of video quality than average viewers are [5]. An experiment using broadcasters as subjects resulted in a model whose coefficients increased, causing a lower estimated MOS. It may be that subjective quality models are also application-dependent. More investigations of this type are needed.

6. Summary

The results of this audiovisual subjective experiment have allowed us to gain insight into how audio and video quality relate to audiovisual quality. For this experiment, video quality was the main component of the overall audiovisual quality. We also found that when the video-only session or the audiovisual session was the third of three sessions, subjects rated the material higher, (by about 0.2 to 0.3 quality units on a 5-point scale) than when the same material appears in the first or second session.

A model of subjective audiovisual quality was also presented. The correlation of this model with the subjective data is 0.978. The model gives significant weight to the video term, reflecting the high correlation coefficient ($\rho = 0.97$) between the video session and the audiovisual session. This model is not likely to be a general model for relating audio and video scores to overall audiovisual scores. However, this model may be extensible to other video teleconferencing data.

As discussed in Section 4, the delay between audio and video can be a significant factor in the overall subjective audiovisual quality. It may be possible to include a measurement of

audio-video differential delay [6] as a factor in the subjective audiovisual model. This would make the model more general, including cases where it is impossible to adjust for the audio-video delay. More work in this area is anticipated.

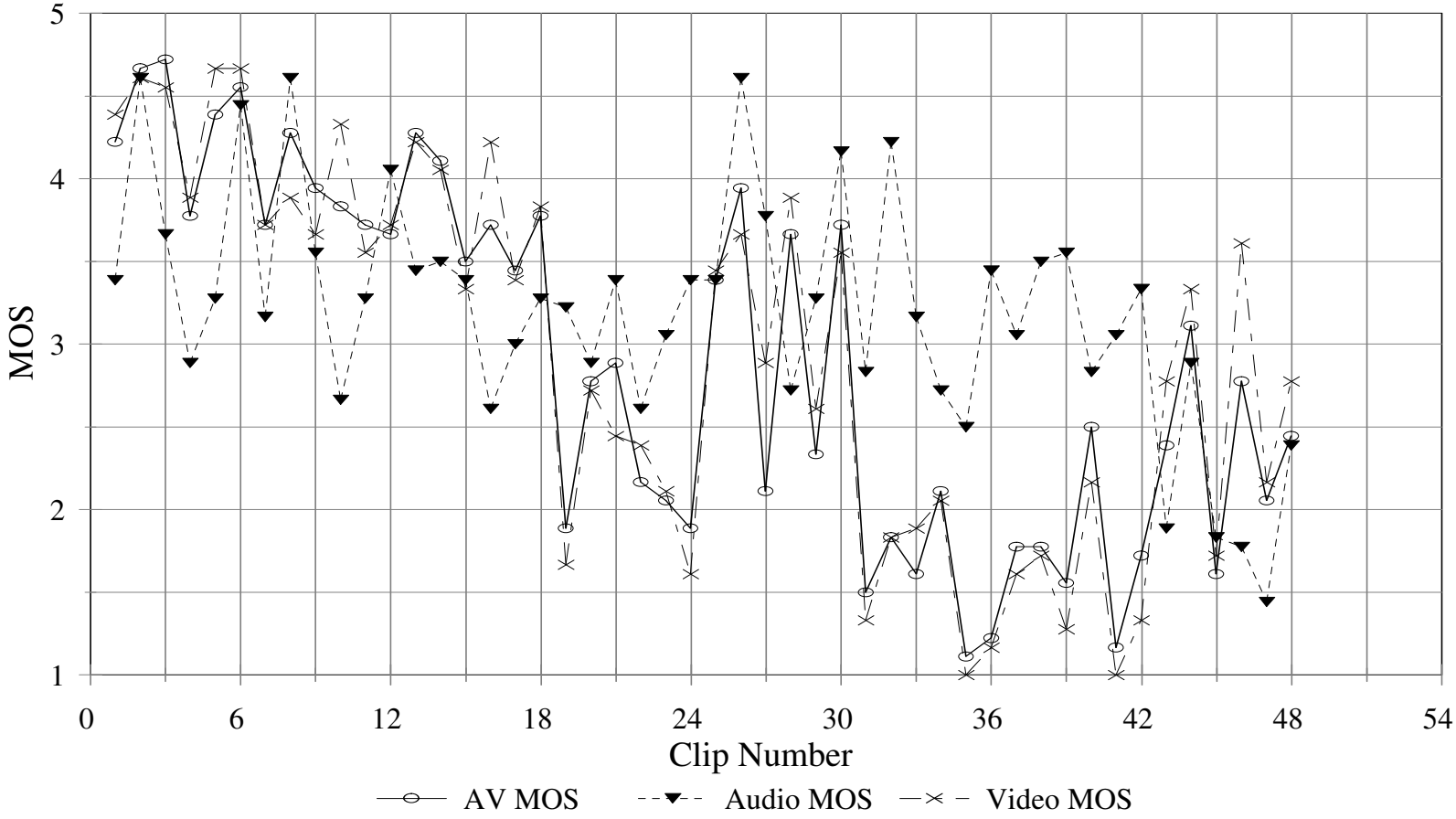

Figure 1: Clip mean opinion score for audio-only test, video-only test, and audiovisual test.
(See Table 2 to relate clip number to processing configuration and scene.)

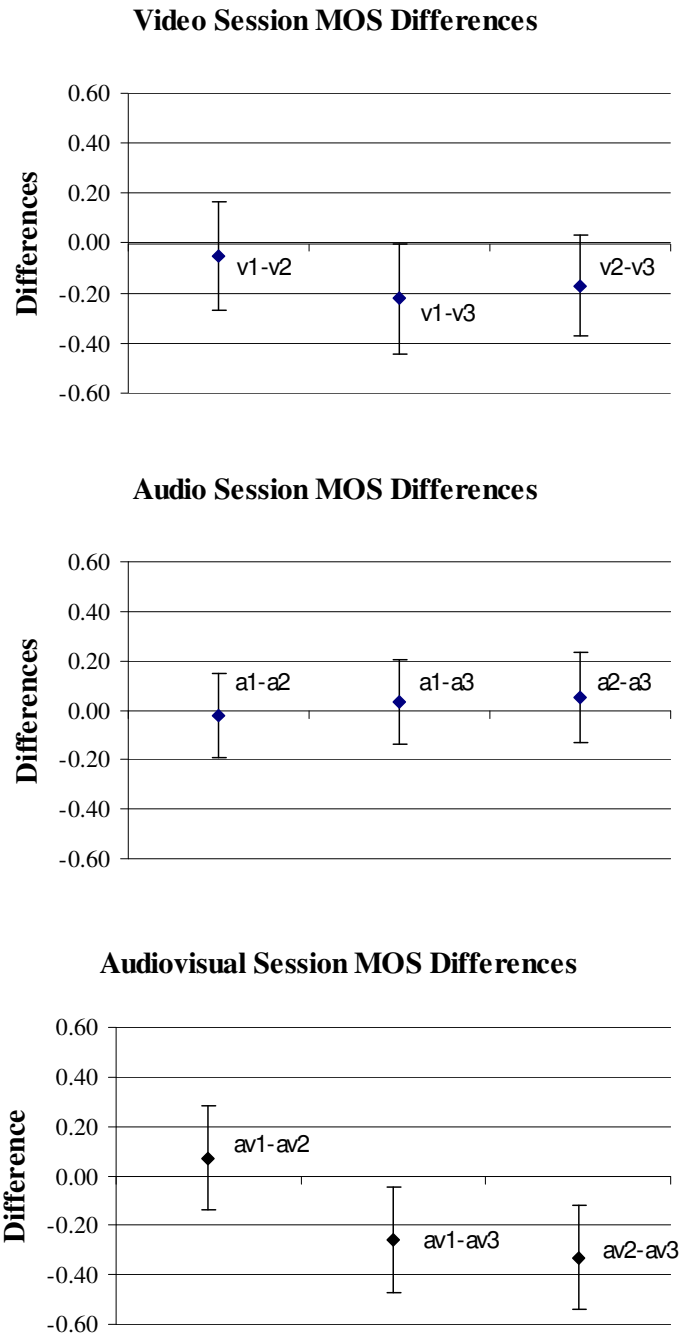

Figure 2: Session MOS differences. The audio session mean opinion scores show no significant differences between session presentation order. However, when the video or audiovisual sessions are presented third, subjects rate the material differently by about 0.2 to 0.3 quality units (on a 5-point scale).

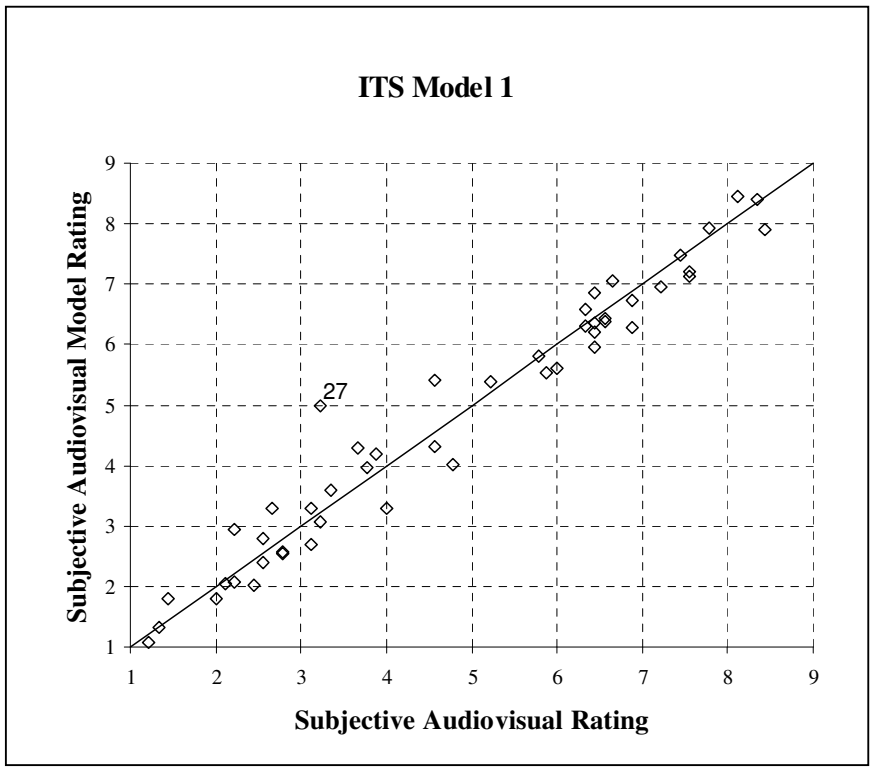

Figure 3: ITS Model 1, $\hat{s}_{av} = -0.677 + 0.217 s_a + 0.888 s_v$, $\rho = 0.978$.

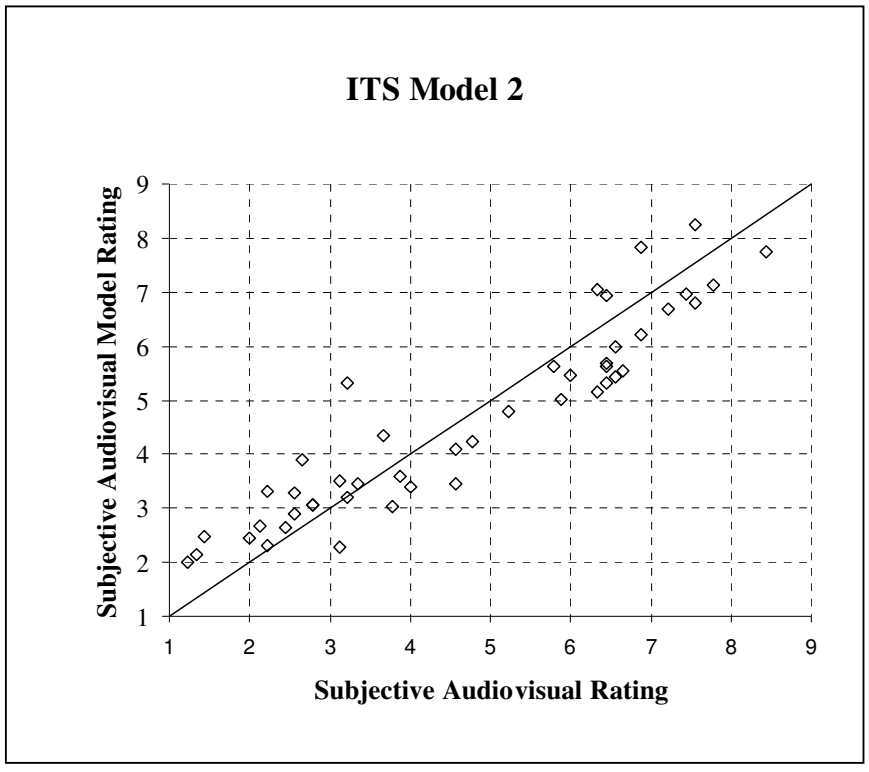

Figure 4: ITS Model 2, $\hat{s}_{av} = 1.514 + 0.121(s_v \times s_a)$, $\rho = 0.927$.

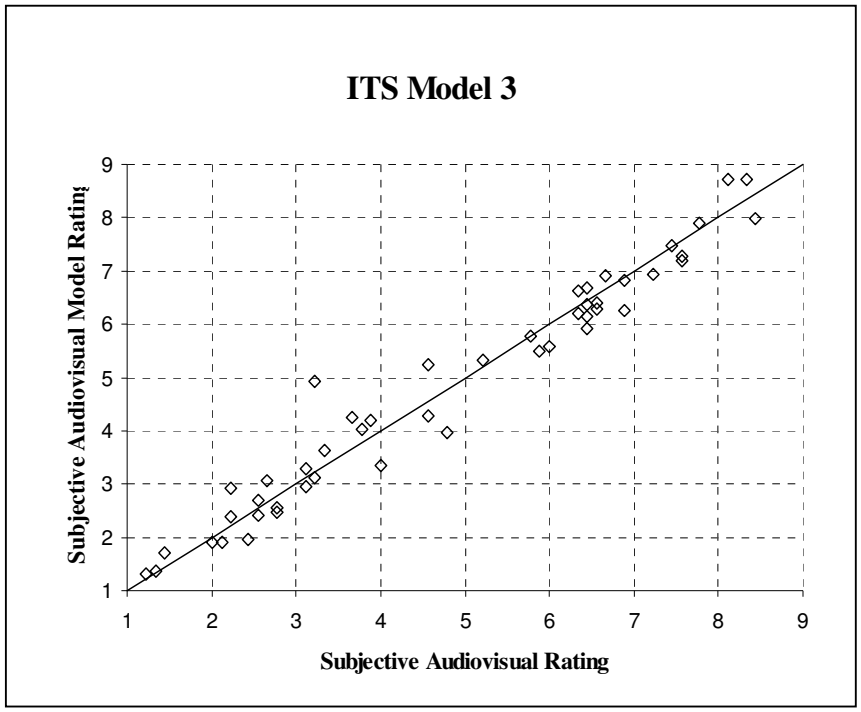

Figure 5: ITS Model 3, $\hat{s}_{av} = 0.517 - 0.0058 s_a + 0.654 s_v + 0.042(s_a \times s_v)$, $\rho = 0.980$.

7. References

[1] ITU-T Recommendation[2] P.910, “Subjective video quality assessment methods for multimedia applications”, Recommendations of the ITU (Telecommunication Standardization Sector).

[2] ANSI T1.801.01-1995, “Digital Transport of Video Teleconferencing-Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment”.

[3] ITU-T Contribution COM 12-19-E, “Relations between audio, video and audiovisual quality”, February 1998, KPN Research, Netherlands.

[4] ANSI-Accredited Committee T1 Contribution[3], T1A1.5/94-124, “Combined A/V model with multiple audio and video impairments”, April 19, 1995, Bellcore, USA.

[5] ANSI-Accredited Committee T1 Contribution, T1A1.5/93-60, “Objective Performance Parameters for NTSC Video at the DS3 Rate”, April 28, 1993, NTIA/ITS.

[6] ITU-T Contribution COM 12-29-E, “Draft new Recommendation on multimedia communication delay, synchronization, and frame rate measurement”, February 1998.

*********end of document**********

[2] Information on obtaining ITU Recommendations and contributions can be found on the ITU’s web page at www.itu.int.

[3] Information on obtaining Committee T1 contributions can be found on T1’s web page at www.t1.org.