Development of Opinion-Based Audiovisual Quality Models for Desktop Video-Teleconferencing

Coleen Jones, D.J. Atkinson
U.S. Department of Commerce/
National Telecommunications and Information Administration/
Institute for Telecommunication Sciences

Abstract: This paper discusses the analysis of an audiovisual desktop video-teleconferencing subjective experiment conducted at the Institute for Telecommunication Sciences. Objective models of the individual audio and video quality are presented. Also discussed is an objective model of the audiovisual quality based upon the results of the individual objective audio and video quality models. Finally, a subjective model of audiovisual quality based upon users' ratings of the audio and video quality is discussed.

1. Introduction

The Institute for Telecommunication Sciences (ITS) conducted an audiovisual desktop video-teleconferencing subjective experiment to investigate the relationship between individual audio and video quality and overall audiovisual quality.

We developed a subjective audiovisual quality model that relates users' ratings of the audio and video quality to their ratings of the audiovisual quality. We also compared our results with subjective audiovisual models developed by other laboratories. Likewise, using our objective measurements of audio and video quality, we have investigated an initial objective audiovisual quality model.

A description of the subjective audiovisual experiment and the results of the subjective and objective data analysis are discussed. The results of this experiment and the models developed may be useful to people developing or using multimedia applications.

2. Subjective Test Plan

The primary goal of this test was to collect 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 was conducted in an acoustically isolated chamber. The audio and video sequences were processed through several representative DVTC configurations. Eighteen subjects were randomly chosen from Department of Commerce staff at the Boulder campus.

This test consisted of three individual sessions presented consecutively with a 10-minute break between 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; and
- an audiovisual-session, in which subjects rated the overall quality of an audiovisual clip.

Subjects were presented all three sessions in one of six possible session orderings, resulting in each of the six session orderings being rated by three subjects.

A source tape in professional ½"-component video format was used as input to each of the eight processing configurations listed in Table 1. Both the input and output of the processing configurations were composite (NTSC) video, since this is the likely format to be used by DVTC users. Because we wanted to remove delay between the audio and video signals as a factor in the audiovisual quality rating, the audio signal was delayed such that the audio and video signals were synchronized. The adjusted audio 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. A personal computer (PC) overlay card was used to display the video clips in SVGA (800x600 pixel resolution) format on a 17" 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]. In an ACR experiment, subjects are presented a single clip that may or may not be degraded. They are then asked to rate the single-ended (i.e., no reference given) quality on a 5-point scale, where 5=excellent, 4=good, 3=fair, 2=poor, and 1=bad.

These numeric scores were then averaged to obtain a mean opinion score (MOS). Averaging over all subjects (18) for each scene-processing configuration combination yields what we term a clip MOS. The clip MOS represents the subjects' average opinion for that specific combination of scene and processing configuration.

The six scenes selected were representative of video-teleconferencing (VTC) scenes [2]. The scenes *vtc1nw* (video-teleconference, clip 1/news announcer) and *smity2* (man called Smity, clip 2) consist of one person (*vtc1nw* has very little motion, and *smity2* has a moderate amount of motion). The scene *vtc2* (video-teleconference, clip 2) has one person with graphics (a map). In the first part of this scene, the woman is talking, and in the second part, there is a camera zoom that causes the whole frame to be in motion. The scene *5row1* (five people in a row, clip 1) has five people sitting around a conference table. And *filter* (block diagram of a digital filter) and *washdc* (map of Washington D.C.) are two graphics-related scenes. Each of the six scenes was processed by all eight processing configurations, resulting in 48 clips that were presented to each subject in each session. Table 2 relates clip number to scene and processing configuration.

Table 1: Processing Configurations

#	Aggregate Bit Rate, System	Video Coding	Audio Coding	Audio Delay (ms)
1	Analog/NTSC[1]	Analog	Analog	0
2	1536 kb/s, A	H.261-C[2]	G.722[3]	80
3	1536 kb/s, B	P[4]	G.722	16
4	384 kb/s, B	H.261 Q[5]	G.711[6]	100
5	384 kb/s, A	H.261 C	G.722	120
6	128 kb/s, A	H.261 Q	G.728[7]	200
7	128 kb/s, B	H.261 Q	G.711	144
8	128 kb/s, B	P	P (8 kb/s)	30

[1] NTSC is 525 line, interlaced, composite video format

[2] H.261 in full common intermediate format (CIF) mode (352x288 pixels)

[3] G.722: 7-kHz bandwidth at 64 kb/s.

[4] Proprietary Coding Algorithm

[5] H.261 in quarter CIF (QCIF) mode (176x144 pixels)

[6] G.711: 4-kHz bandwidth at 64 kb/s.

[7] G.728: 4-kHz bandwidth at 16 kb/s.

Table 2: Determination of Clip Number, from Configuration Number and Scene Name.

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

3. Subjective Results

The clip MOSs for all 48 clips (6 scenes and 8 processing configurations) are shown in Figure 1. The clip MOSs are plotted for the audio-only session (▼), the video-only session (×), and the audiovisual session (○).

It is interesting to note the difference between the video MOSs for the scene *vtc1nw* for the first three configurations, i.e., the NTSC configuration (clip 4) and the two 1536-kb/s configurations (clips 10 and 16). The MOS for clip 4 was 3.89, and the MOSs for clips 10 and 16 were 4.33 and 4.22 respectively. One would expect that the NTSC video scene 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 clips on a PC monitor. The overlay card uses an 8-bit color palette to display video on the PC monitor. In the NTSC video scene, 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 scene was fed through the overlay card for display, it exhibited poor color 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 system), the audio MOS varied by more than 1½ quality units (see Figure 1), which is larger than would normally be expected. The other processing configurations exhibited this pattern as well, as seen in Figure 1. 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 tracks (with background noise).

The confidence intervals on the video (0.293 average) and audiovisual (0.338 average) MOSs are reasonable. However, the confidence intervals on the audio (0.373 average) MOSs are larger than typically found in ACR audio tests (0.2 to 0.25). This is most likely due to the variation in source audio quality as discussed above.

It appears that video quality was the main factor in audiovisual quality for the systems tested (see Table 3). These results are similar to those obtained by KPN Research [3] in a similar experiment that resulted in correlation coefficients of $\rho_{a,v} = -0.02$, $\rho_{a,av} = 0.33$, and $\rho_{v,av} = 0.90$.

Table 3: Between-test Correlation Coefficients for the 48 Clip Mean Opinion Scores

Test Comparison	ρ
audio and video sessions $(\rho_{a,v})$	0.29
audio and audiovisual sessions $(\rho_{a,av})$	0.41
video and audiovisual sessions $(\rho_{v,av})$	0.97

We conducted an analysis to determine whether or not the session ordering was significant. We calculated the session MOS by averaging over all 48 clips and all subjects who saw a given session either first, second, or third during their testing. For example, 6 subjects rated video in the third session (audio, audiovisual, video; or audiovisual, audio, video). Thus, the session MOS for subjects who rated video third ($\bar{v}_3$) was averaged over these 6 subjects and all 48 test clips. We then calculated the session MOS differences (three differences each for the video

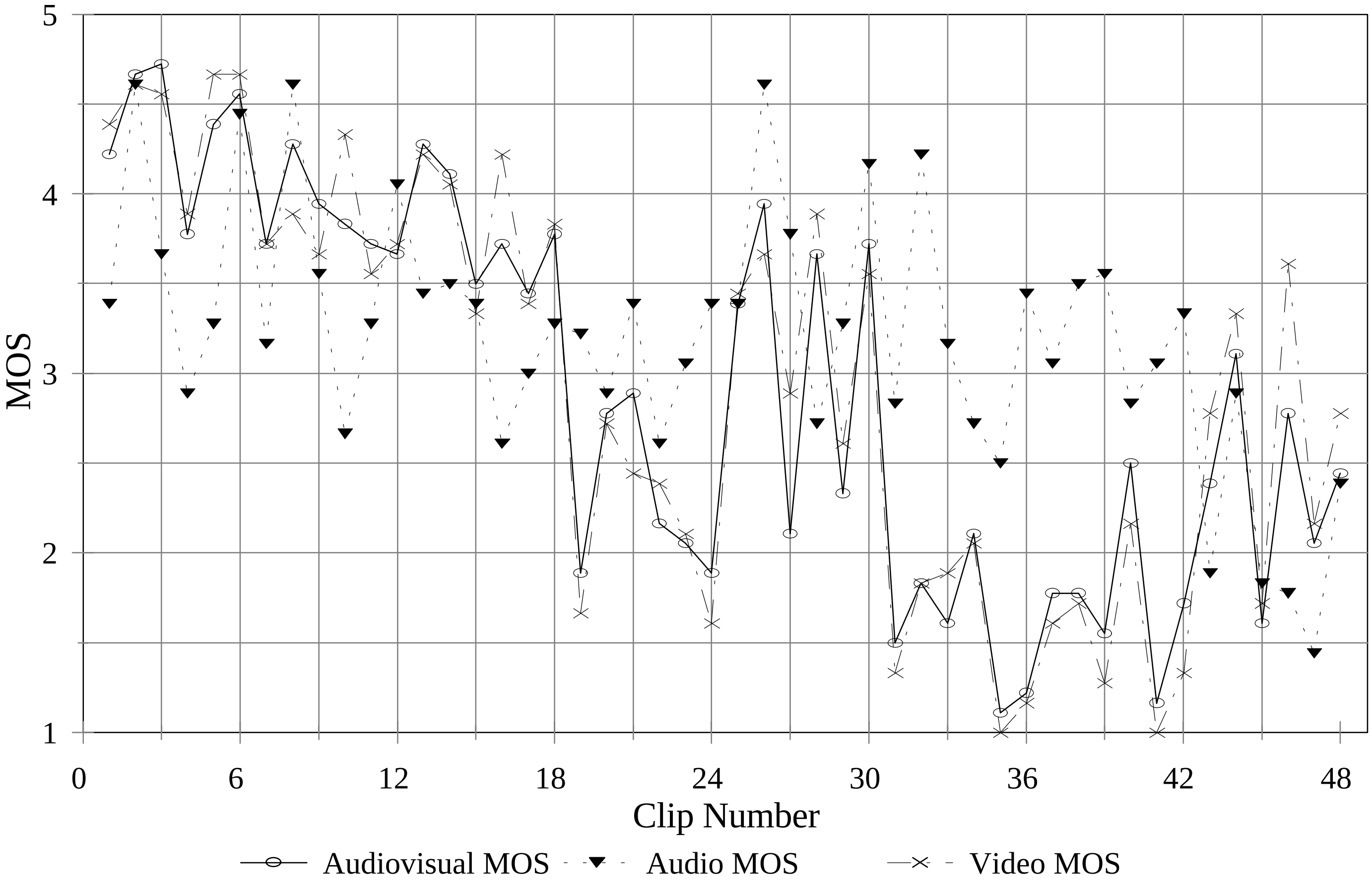

Figure 1: Clip mean opinion score for audio-only test, video-only test, and audiovisual test. (See Table 2 to relate clip number to processing configuration and scene.)

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 did not span zero, that difference was deemed significant. Table 4 lists the session MOS differences and confidence intervals (CIs). The CIs assume an approximate Gaussian distribution given the large number of samples over which we averaged.

Table 4: Session MOS Differences

Comparisons	Session MOS Difference	Half-width CI (95%)	Confidence Interval Bounds	
$\bar{v}_1 - \bar{v}_2$	-0.051	0.217	-0.268	0.166
$\bar{v}_1 - \bar{v}_3$	-0.222	0.218	-0.44	-0.004
$\bar{v}_2 - \bar{v}_3$	-0.172	0.202	-0.374	0.03
$\bar{a}_1 - \bar{a}_2$	-0.021	0.173	-0.194	0.152
$\bar{a}_1 - \bar{a}_3$	0.031	0.170	-0.139	0.201
$\bar{a}_2 - \bar{a}_3$	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

Table 4 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, for the video and audiovisual sessions, subjects rated video and audiovisual quality higher (by about 0.2 to 0.3 quality units) in the third session than the first two sessions. The video session MOS differences are near zero at the bounds of the confidence interval. However, the differences are not as small as one would expect. 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 needed to minimize them.

4. Objective Results

Over the last several years, the staff at ITS has developed perception-based measurements that objectively quantify distortions caused by digital compression. Measurements have been developed individually for audio quality and video quality. These measurements can be combined into models that predict users' opinions of quality. The coefficients in the models are determined by fitting the objective measurements to a set of subjective data. Objective model results are compared with subjective data to determine the model's performance. This subjective experiment is the first step in an attempt to develop a combined objective audiovisual quality measurement. The objective video quality and audio quality models are described below, followed by our initial attempt at objectively modeling the overall audiovisual quality results.

4.1 Objective Video Quality Model

ITS-developed objective video quality parameters (metrics) are part of ANSI T1.801.03 [4], and the objective video quality model was designed using the ANSI T1.801.03 parameters. We opted to use a two-parameter model, because a two-parameter model gives a 24:1 clip to parameter ratio.[1] The two-parameter model that best correlated with the video session clip MOSs was selected. The objective video model, denoted $\hat{s}_v$, is

$$\hat{s}_v = 4.679 - 1.272P_{714} - 3.770P_{718} \Big|_1 . \quad (1)$$

The notation $\Big|_1$ indicates clipping the estimated MOS at 1 when it is less than 1. Parameter P_{714} is the average lost motion energy with noise removed. This parameter is defined in section 7.1.4 of ANSI T1.801.03. It essentially measures the amount of lost motion or jerkiness in the output video relative to the input video. Parameter P_{718} is the average edge energy difference. This parameter is defined in section 7.1.8 of ANSI T1.801.03. This parameter can quantify lost edges (e.g. blurring) or added edges (e.g. blocking) individually, but when both blurring and blocking are present, it does not effectively measure the blocking artifacts. For this particular data set, the blurring degradation overwhelmed the blocking artifacts. Therefore, P_{718} is a measure of the amount of blurring in the output video relative to the input video. The correlation coefficient between the between the subjective video clip MOSs and the model in equation

---

[1] Our guideline is to use a minimum data-to-parameter ratio of at least 20:1 to avoid over-modeling the data.

(1) is 0.925 (ρ^2 = 0.856), explaining 86% of the variance in the subjective data. The scatter plot of the subjective video clip MOSs versus the video model output is shown in Figure 2.

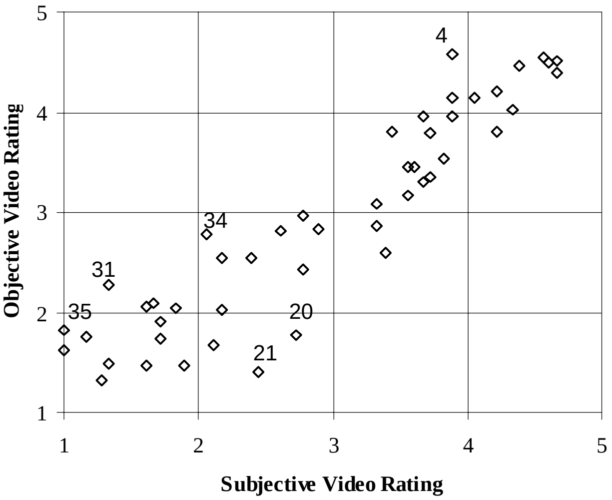

Figure 2: Results of the objective video quality model (numbered clips are referenced in text).

Several factors account for the outliers produced by this model. In cases where the error is negative (i.e., the objective model produced scores that were too high, for example clips 31, 34, and 35 in Figure 2.), the model did not fully quantify all of the degradations within the video scene. This is mainly due to the fact that this two-parameter model does not effectively measure blocking artifacts in the presence of blurring. When a third parameter is added to detect blocking, these errors decrease. One exception is clip 4, the NTSC configuration. Of course blocking is not a factor in an NTSC configuration. The previously mentioned interaction between this NTSC-scene combination and the overlay card caused the subjects to lower their ratings. On the other hand, the video measurement system bypassed the overlay card and used the viewing tape directly, which increased the difference between the objective model output and the subjective rating.

For two outliers, the error was positive (i.e., the objective model produced scores that were too low). In clip 21, the lines on the man's shirt were blurred by the system. This blurring was not objectionable to viewers in this ACR test, because viewers did not care that the shirt was blurred or were unaware of how the shirt should look. Thus, the model rated the clip lower than the viewers did. Clip 20 contains a block diagram of a filter and exhibits a distinct ordering effect. When subjects rated this clip in the first session, its MOS was 4.4, but in sessions two and three, the MOS was about 2. This may be due to the ordering of clips on the first tape, or perhaps subjects became more familiar with higher quality versions of this clip in the second and third sessions.

4.2 Objective Audio Quality Model

The objective audio quality model was designed using the ITS-developed Measuring Normalizing Block algorithm, structure 1 (MNB1) [5,6] with a bandwidth-compensation factor. The MNB1 algorithm measures the quality of narrowband voice by transforming the input and output audio signals into a perceptual domain. The perceptually transformed signals are then compared using the MNBs to detect frequency and temporal distortions in the output relative to the input. The output of this algorithm is auditory distance (AD). It is a measure of how different the output audio signal is from the input audio signal. Thus, larger auditory distances indicate poorer output audio quality. Auditory distance was linearly fit to the audio session subjective data. The objective audio model, denoted $\hat{s}_a$, is

$$\hat{s}_a = 4.388 - 0.638AD . \quad (2)$$

As shown in Table 1, some of the configurations tested used the G.722 audio coding algorithm. This was unavoidable because we did not have our choice of audio coding algorithm for some configurations. G.722 is a wideband audio coding algorithm. It codes 7-kHz audio at 64 kb/s. To measure the quality of this coding algorithm, an objective audio quality metric that analyzes the 7-kHz bandwidth is necessary. Because a wideband quality metric has not yet been developed, we experimentally applied our narrowband (4 kHz) objective audio quality measurement (MNB1) with a bandwidth-compensation factor. This experimental measurement technique did not perform well under these conditions. However, the MNB1 algorithm has performed quite well in narrowband (4 kHz) experiments over a broad range of coding algorithms [6].

Figure 3 contains a scatter plot of the subjective audio clip MOSs versus the audio model output. The interaction between the wideband codec and the noise in the source scene is best observed in configurations 1 (NTSC), 2 and 5 (both G.722). For the scenes with noisy audio (*5row1*, *smity2*, *vtc1nw*, *vtc2*), the audio quality was rated higher by the objective model than by the subjects. This is because the objective metric quantifies differences between the input and output signals, whereas the subjects only hear the output audio clip. Because the subjects only hear the output audio clip, they consider noise as an impairment relative to their inherent reference (see clips 10, 16, 22, 28, and 45 in Figure 3). For the scenes with clean audio, the objective metric was unable to adequately assess the wideband nature of the signal, and thus the audio model rated the audio lower than the subjects did (see clips 6, 8, 12, 26, and 32 in Figure 3). The objective audio quality model results could be improved with the development of a wideband measurement, and the use of degradation category rating (DCR) subjective tests where subjects hear both the input and output audio clip.

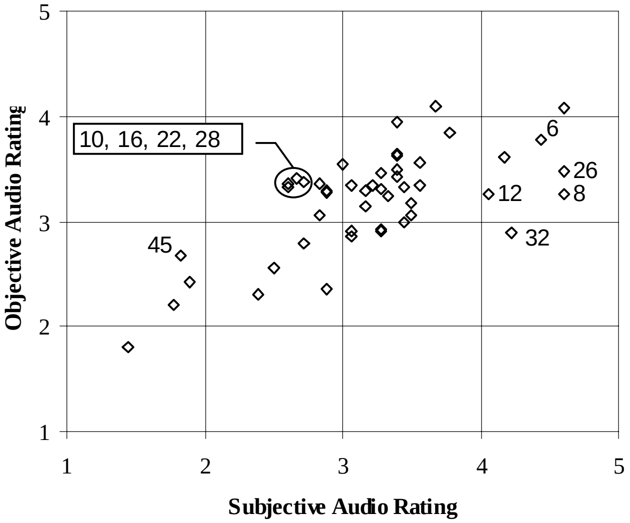

Figure 3: Results of the objective audio quality model (numbered clips are referenced in the text).

4.3 Objective Audiovisual Model

As a simple initial attempt, the objective audiovisual quality model ($\hat{s}_{av}$) was built using the output of the individual objective audio and video models discussed above (equations (1) and (2)). Several different forms of equations were analyzed including cross products (between the audio and video model outputs), and the sums and differences of first- and second-order terms in different permutations. The model that correlated best with the audiovisual subjective data was a simple linear combination of the output of the individual audio and video objective models.

$$\hat{s}_{av} = -0.949 + 0.854\hat{s}_v + 0.422\hat{s}_a \qquad (3)$$

The correlation coefficient between the subjective audiovisual scores and the model in equation (3) is 0.91 ($\rho^2 = 0.827$). The scatter plot of the subjective audiovisual clip MOSs versus the objective audiovisual model output is shown in Figure 4.

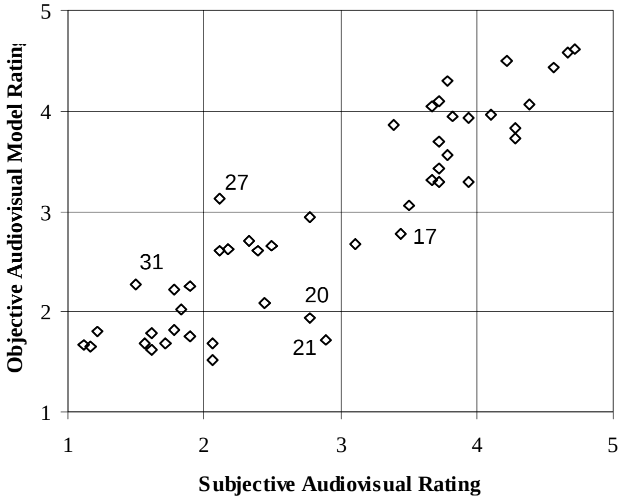

Figure 4: Results of the objective audiovisual quality model (numbered clips are referenced in the text).

As seen in Table 3, the subjective data shows that for this test, video quality is the primary factor in the overall audiovisual subjective score. Thus, it is not surprising that the audiovisual model has difficulty with some of the same clips as does the video model (e.g., clips 17, 20, 21, and 31 in Figure 4). Clip 27 is discussed in Section 5. A more accurate objective audiovisual model might be developed if better results could be obtained from the objective audio quality model. Other mathematical relationships between audio quality, video quality, and audiovisual quality may also prove to be useful for improving the audiovisual model.

5. Subjective Audiovisual Models

Finally, a model that relates the individual subjective audio and video MOSs (s_a, s_v) to the subjective audiovisual MOSs was investigated. As with the objective audiovisual model, several different forms of equations were analyzed. Two models had similar correlation coefficients (approximately 0.98) when compared with the audiovisual subjective data. The addition of a cross term in one of the models did not significantly improve the correlation coefficient; therefore we used the simpler linear combination:

$$\hat{s}_{av} = -0.677 + 0.888 s_v + 0.217 s_a \qquad (4)$$

The correlation coefficient between the subjective audiovisual scores and the model in equation (4) is 0.978 ($\rho^2 = 0.957$). The scatter plot of the subjective audiovisual clip MOSs versus the subjective audiovisual model output is shown in Figure 5.

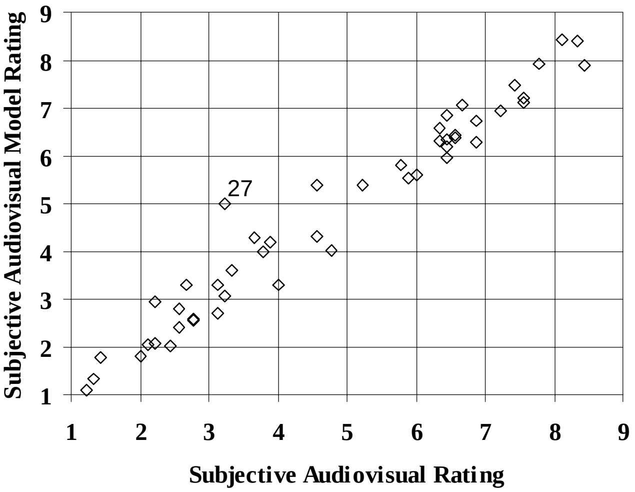

Figure 5: Results of the subjective audiovisual quality model (numbered clips are referenced in the text).

For the subjective model, all of the subjective data was converted to a 9-point scale so comparisons could be made with subjective models from other laboratories that used a 9-point ACR scale. We typically consider prediction errors greater than ½ quality unit to be significant. However, because we have converted this data to a 9-point scale, prediction errors greater than 1 quality unit are significant. All of the prediction errors for this model were less than one quality unit, with the exception of clip 27 that exhibited a relatively large error of –1.77. Because we did not want delay between the audio and video signals to be a factor in this experiment, we delayed the audio signal so that it would be synchronized with the video signal. We chose a fixed audio 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 signals 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. It may be possible to include a measurement of audio-video differential delay [7] 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.

Other laboratories have conducted similar experiments and developed subjective audiovisual models [3,8]. Table 5 summarizes results from ITS, KPN Research, and Bellcore. All three laboratories investigated a model based upon the product of the individual audio and video subjective scores from one or more experiments. All three laboratories 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 Research model 1 and Bellcore models 1 and 2. This may be due to either the noisy source material, or the different impairments used in our experiment.

KPN Research model 2 adds the product term to the linear model. KPN Research 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 Research 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 factor is even negative which is counter-intuitive, and should be set to zero.

With our objective models, developed in other experiments, 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 are average viewers. An experiment using broadcasters as subjects resulted in a model whose coefficients increased, causing a lower estimated MOS [9]. It may be that subjective quality models are also application-dependent. More investigations of this type would be interesting.

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 than when the same material appears in the first or second session.

The objective data analysis showed that the objective video and audiovisual models predicted subjective results acceptably. The audio metric (modified to attempt to account for the wide bandwidth nature of the audio) did not perform as well as hoped. In the future we may design an audiovisual experiment more specific to the task of relating audio and video quality to audiovisual quality. Using high-quality source audio and systems and data rates that fall within the scope of our audio metric, we should be able to improve both the results of the objective audio model and the objective audiovisual model as well.

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[3], "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 T1.801.03-1996, "Digital Transport of One-Way Video Signals – Parameters for Objective Performance Assessment".

[5] ANSI-Accredited Committee T1 Contribution[4], T1A1.7/97-003R1, "Additional Information on Proposed Objective Quality Measure for the Audio Portion of an Audio-Visual Session", April 30, 1997, NTIA/ITS, USA.

[6] Voran S. "Estimation of perceived speech quality using measuring normalizing blocks." Proceedings of the 1997 IEEE Speech Coding Workshop, Pocono Manor PA, USA, 1997, pages 83-84.

[7] ANSI T1801.04-1997, "Multimedia Communications Delay, Synchronization, and Frame Rate".

[8] ANSI-Accredited Committee T1 Contribution, T1A1.5/94-124, "Combined A/V model with Multiple Audio and Video Impairments", April 19, 1995, Bellcore, USA.

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

Table 5: Comparison of Subjective Audiovisual 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 Research	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

---

[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 ANSI Standards can be found on ANSI's web page at www.ansi.org.

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