

Relating Audio and Video Quality Using CIF Video

Mark A. McFarland
Margaret Pinson
Carolyn Ford
Arthur Webster
William Ingram
Scott Hanes
Kelsey Anderson



technical memorandum

Relating Audio and Video Quality Using CIF Video

**Mark A. McFarland
Margaret Pinson
Carolyn Ford
William Ingram
Scott Hanes
Arthur Webster
Kelsey Anderson**



U.S. DEPARTMENT OF COMMERCE

September 2010

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendations or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

CONTENTS

Page

| | |
|--|-----|
| FIGURES | vi |
| TABLES | vi |
| ABBREVIATIONS/ACRONYMS..... | vii |
| 1 INTRODUCTION | 1 |
| 2 SUBJECTIVE TEST DESIGN..... | 7 |
| 2.1 Source Sequences | 7 |
| 2.2 Subjective Test Method..... | 10 |
| 3 DATA ANALYSIS..... | 11 |
| 3.1 Audio and Video Rated Together (Audiovisual Ratings) | 11 |
| 3.2 Audio and Video Rated Separately | 15 |
| 4 CONCLUSIONS..... | 20 |
| 5 FUTURE WORK..... | 20 |
| 6 REFERENCES | 21 |

FIGURES

| | Page |
|--|------|
| Figure 1. Basic components of a multimedia objective quality model..... | 2 |
| Figure 2. Graphical representation of coefficient values for models containing intercept, s_a , and s_v terms (green models)..... | 4 |
| Figure 3. Graphical representation of coefficient values for models containing intercept and $(s_v \times s_a)$ term (blue models)..... | 5 |
| Figure 4. Graphical representation of coefficient values for models containing intercept, s_a , s_v , and $(s_v \times s_a)$ terms (red models). | 6 |
| Figure 5. Mean audiovisual MOS when audio was at highest quality, averaged for each video HRC. This indicates the quality of each video HRC, influenced by the original audio quality..... | 12 |
| Figure 6. Mean audiovisual MOS when video was at highest quality, averaged for each audio HRC. This indicates the quality of each audio HRC, influenced by the original video quality..... | 13 |
| Figure 7. Audiovisual MOS predicted by weighted video MOS (top), and errors (bottom)..... | 16 |
| Figure 8. Audiovisual MOS predicted by weighted audio MOS (top), and errors (bottom)..... | 17 |
| Figure 9. Audiovisual MOS predicted by audio \times video MOS (top), and errors (bottom). | 18 |
| Figure 10. Audiovisual MOS predicted by weighted audio, video, and audio \times video MOS (top), and errors (bottom). | 19 |

TABLES

| | Page |
|---|------|
| Table 1. Comparison of Subjective Audiovisual Models from Different Laboratories (like models grouped by color) | 3 |
| Table 2. Example Video Frame from each Source Clip | 7 |
| Table 3. Audio-Video Sequence Descriptions and Pools | 8 |
| Table 4. Video HRCs..... | 9 |
| Table 5. Audio HRCs..... | 9 |
| Table 6. ANOVA of individual audio-video clips..... | 15 |
| Table 7. Linear combination equations, ρ , and ρ^2 of experiment. | 19 |

ABBREVIATIONS/ACRONYMS

| | |
|--------------|---|
| ACR | Absolute Category Rating |
| CIF | Common Intermediate Format (352 pixels by 288 lines) |
| DOC | Department of Commerce |
| FPS | frames per second |
| GUI | graphical user interface |
| HDTV | High-Definition Television |
| ITS | Institute for Telecommunication Sciences |
| HRC | Hypothetical Reference Circuit |
| SG9 | ITU-T Study Group 9 |
| ITU-R | International Telecommunications Union, Radiocommunication Standardization Sector |
| ITU-T | International Telecommunications Union, Telecommunication Standardization Sector |
| MOS | mean opinion score |
| MPEG | Moving Picture Experts Group |
| NTIA | National Telecommunications and Information Administration |
| NTSC | National Television System Committee |
| PC | personal computer |

RELATING AUDIO AND VIDEO QUALITY USING CIF VIDEO

Mark A. McFarland, Margaret Pinson, Carolyn Ford, Arthur Webster, William Ingram, Scott Haines, Kelsey Anderson¹

NTIA/ITS² has conducted a series of studies to quantify the effects that individual audio and video qualities have on the overall Mean Opinion Score (MOS) for a given set of audiovisual clips. The experiment described in this report studies the effects that the synthesis of audio and video quality has on a subject's overall MOS. The overall MOS for this set of audiovisual clips can be predicted rather well from the video MOS alone. This appears to be a consequence of the wide quality range spanned by the video impairments, in combination with the narrow quality range spanned by the audio impairments. This result will not necessarily be valid for other choices of audiovisual material.

Key words: video quality; subjective testing; audio quality multimedia quality

1 INTRODUCTION

Much work has been done to characterize subjective quality of video and audio independent of each other [1], [2]. Work has also been done to answer the question of how audio quality and video quality combine to form a person's opinion of an audiovisual sequence when viewed and heard simultaneously [3]-[6]. We are interested in discovering mathematical functions that describe audiovisual (or "multimedia") quality.³ Specifically, we would like to explore whether audiovisual quality can be described as a function of the audio quality and video quality measured separately. NTIA/ITS is conducting a series of experiments to this end and this report documents the first of these experiments.

The International Telecommunications Union, Telecommunication Standardization Sector (ITU-T), Study Group 9 (SG9), via Recommendation J.148 [7] describes a method for obtaining an objective measure of multimedia quality from separate (objective) measures of the video quality, the audio quality, and the desynchronization of the two signals. Task information can also be considered. In the current experiment, subjective measures of the qualities of audio only, video only, and both audio and video together are used to determine the multimedia quality integration function. Desynchronization is not considered in this experiment.

¹ The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

² The Institute for Telecommunication Sciences (ITS) is the research and engineering branch of the National Telecommunications and Information Administration (NTIA), a part of the U.S. Department of Commerce (DOC).

³ The term "quality" in this context refers to the subjective aesthetic impression reported by a human viewer, and not to the intelligibility (or intelligence value) of the multimedia sample for human or automatic recognition.

Figure 1, reproduced from ITU-T Rec. J.148, shows how audio quality, differential delay, and video quality can be combined in the integration function. The purpose of the multimedia quality integration function (i.e. the multimedia model) is to integrate qualities from the audio and video to predict the overall multimedia quality held to be representative of human perception. The inputs to the integration function are:

1. Aq: a measurement of audio quality calculated by the objective audio model (e.g., ITU-T Rec. P.862 [8])
2. Vq: a measurement of visual quality calculated by the objective video image model (e.g., ITU-T Rec. J.144 [2])
3. Differential delay: a measure of the synchronization error between the audio and video sources (ITU-R Rec. BT.1359-1 [9])
4. Task: the task being performed, which indicates the amount of interactivity associated with the service being examined (e.g., whether the multimedia is intended to be used for videoconferencing, telemedicine, entertainment, etc.).

In the current experiment, subjective scores are used for #1 and #2 rather than objective model outputs. Inputs #3 and #4 are not used.

The multimedia integration function estimates three quantities:

1. Objective measurement of audio quality, accounting for the influence of video quality, labeled $Aq(Vq)$.
2. Objective measurement of video quality, accounting for the influence of audio quality, labeled $Vq(Aq)$.
3. Overall multimedia quality, or the quality of the audiovisual sequence taken as a whole, labeled as “Multimedia quality.”

The current experiment focuses solely upon output #3.

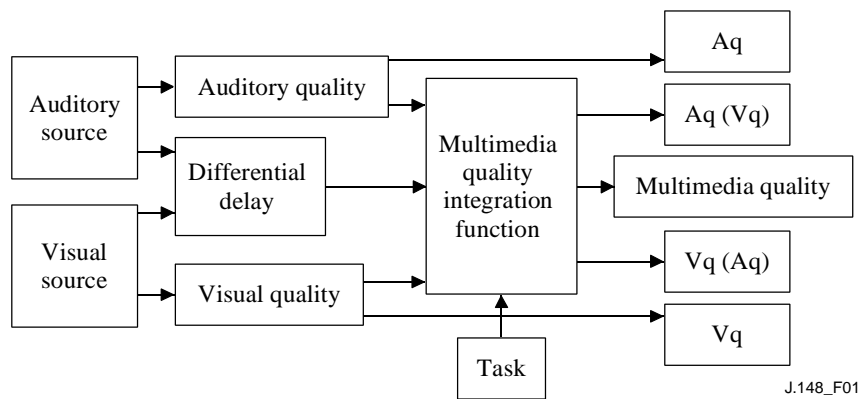


Figure 1. Basic components of a multimedia objective quality model.

As mentioned above, NTIA/ITS is conducting a series of studies that focus on the prediction of overall subjective multimedia quality given the core inputs. The task for these studies is the aesthetic impression (i.e., perceptual quality) of multimedia signals, of varying video resolutions, that have been processed from the original signals (e.g., compressed, subjected to transmission errors, etc.). The current experiment uses CIF video in which the audio and video are synchronized. Future studies will explore other video resolutions as well as desynchronization of the video and audio signals.

NTIA/ITS and other laboratories have performed similar work (e.g., [5], [10]–[12]). The results of that work, along with a summary of the results of this experiment, are presented in Table 1, where similar models are printed in the same color for clarity. In this report, \hat{s}_{av} corresponds to the predicted subjective MOS audiovisual score, s_a corresponds to the subjective audio MOS, s_v corresponds to the subjective video MOS, $(s_v \times s_a)$ is the audio-video cross term, ρ is the correlation coefficient, and ρ^2 is the variance.

Table 1. Comparison of Subjective Audiovisual Models from Different Laboratories (like models grouped by color)

| Laboratory | Model | ρ | ρ^2 |
|----------------------------|--|--------|----------|
| NTIA/ITS (1998) [5] | $\hat{s}_{av} = -0.677 + 0.217s_a + 0.888s_v$ | 0.978 | 0.957 |
| | $\hat{s}_{av} = 1.514 + 0.121(s_v \times s_a)$ | 0.927 | 0.859 |
| | $\hat{s}_{av} = 0.517 - 0.0058s_a + 0.654s_v + 0.042(s_v \times s_a)$ | 0.980 | 0.960 |
| KPN Research [10] | $\hat{s}_{av} = 1.45 + 0.11(s_v \times s_a)$ | 0.97 | 0.94 |
| | $\hat{s}_{av} = 1.12 + 0.007s_a + 0.24s_v + 0.088(s_v \times s_a)$ | 0.98 | 0.96 |
| Bellcore [11] | $\hat{s}_{av} = 1.07 + 0.111(s_v \times s_a)$ | 0.99 | 0.99 |
| | $\hat{s}_{av} = 1.295 + 0.107(s_v \times s_a)$ | 0.99 | 0.98 |
| University of Tsukuba [12] | $\hat{s}_{av} = 0.618 + 0.211s_a + 0.188s_v + 0.112(s_v \times s_a)$ | 0.918 | 0.843 |
| NTIA/ITS (2010) | $\hat{s}_{av} = -0.5875 + 0.3599s_a + 0.8037s_v$ | 0.92 | 0.85 |
| | $\hat{s}_{av} = 1.1096 + 0.1959(s_v \times s_a)$ | 0.93 | 0.86 |
| | $\hat{s}_{av} = 0.7500 - 0.0452s_a + 0.3882s_v + 0.1250(s_v \times s_a)$ | 0.918 | 0.843 |

Figures 2, 3 and 4 show a graphical representation of the coefficients for similar models from each laboratory. Figures 2 and 3 show similar results for the given models. In Figure 4, both ITS models show a negative coefficient for the s_a term, while the models from KPN Research [10] and University of Tsukuba [12] show a positive coefficient for the s_a term. It seems counterintuitive that audiovisual quality increases as audio quality decreases. The reason for the negative s_a coefficient is most likely because s_a is overrepresented in the $(s_v \times s_a)$ term in the ITS models. The Bellcore models [11] had the highest correlations.

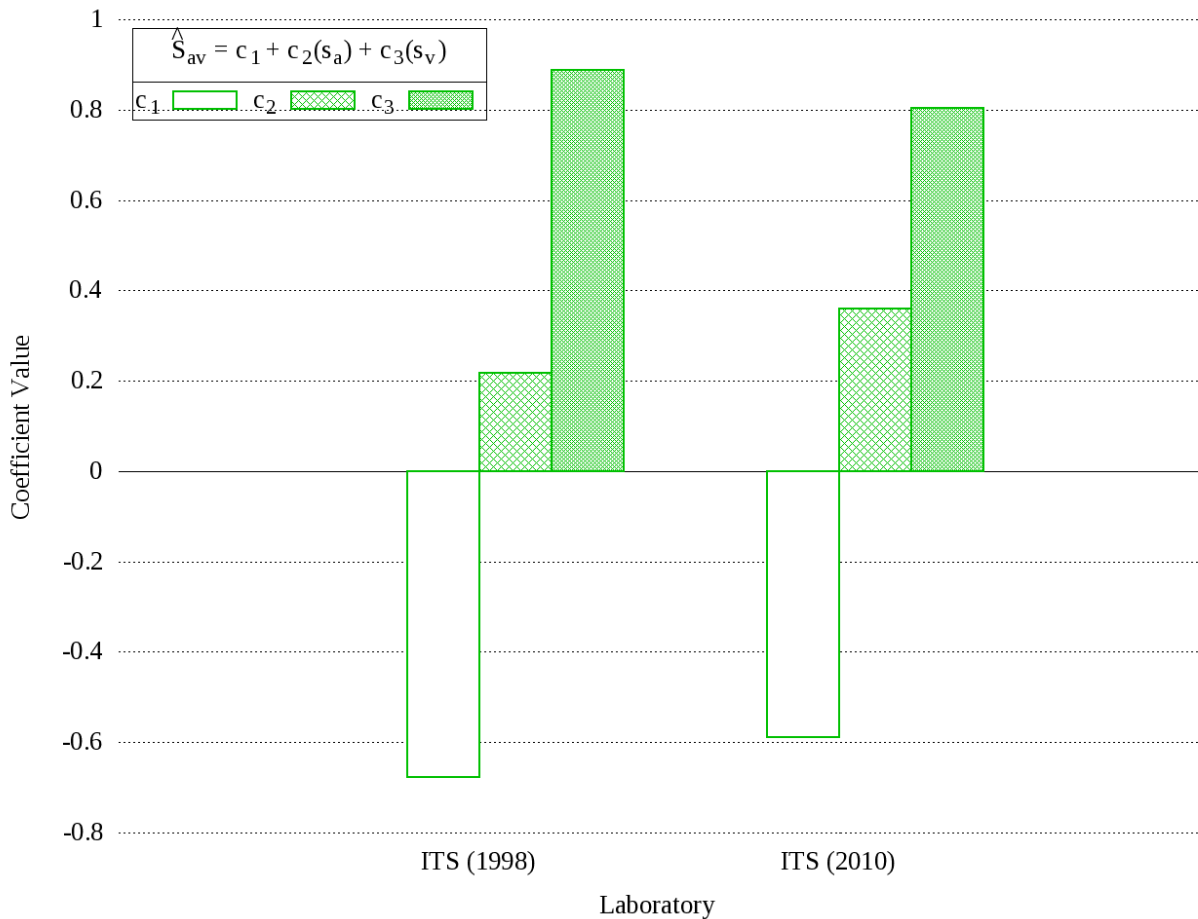


Figure 2. Graphical representation of coefficient values for models containing intercept, s_a , and s_v terms (green models).

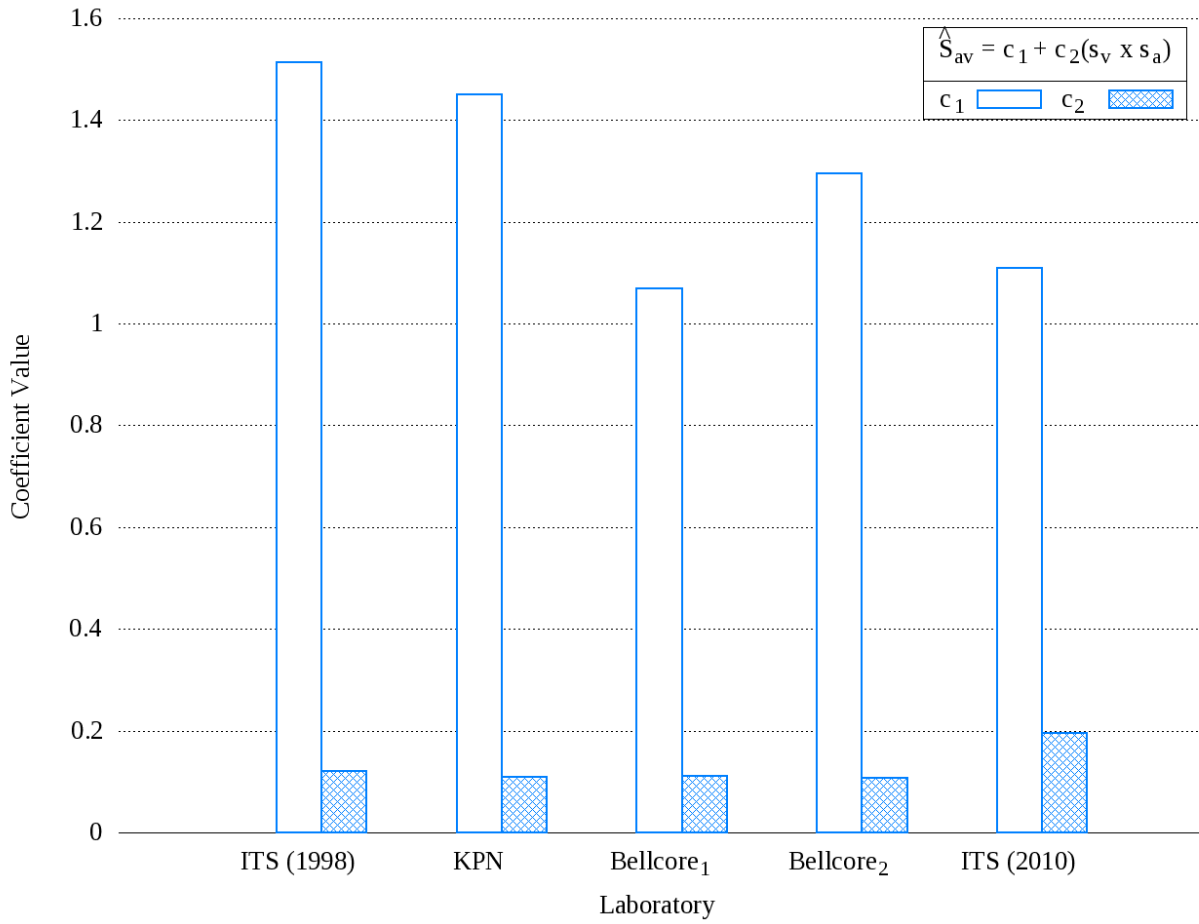


Figure 3. Graphical representation of coefficient values for models containing intercept and $(s_v \times s_a)$ term (blue models).

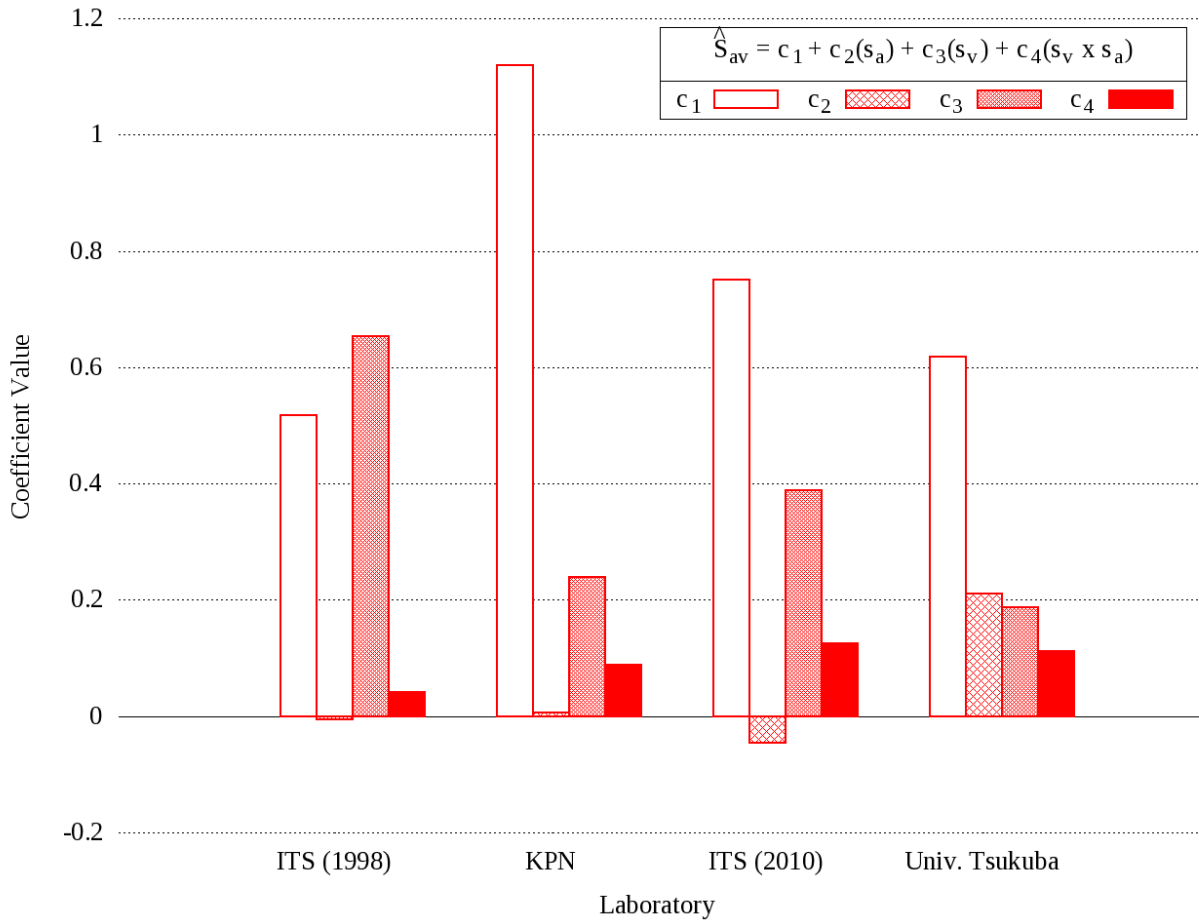


Figure 4. Graphical representation of coefficient values for models containing intercept, s_a , s_v , and $(s_v \times s_a)$ terms (red models).




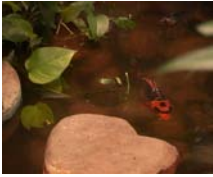





2 SUBJECTIVE TEST DESIGN

2.1 Source Sequences

This experiment examined ten video clips. The quality of the original audio and video recordings varied somewhat, though all had at least good quality as judged by the experiment designers. Of the ten clips, eight were filmed with audio. Slight quality differences were apparent in those audio recordings (e.g., the clip filmed in an automobile contains slight reverberation). The remaining two clips were paired with Harvard Balanced Sentences that had been recorded separately in a sound isolation room, resulting in a very clean audio speech signal. The video quality ranged from a VHS recording, to a professional grade HDTV camcorder. The original video was de-interlaced and scaled from the original resolution (NTSC or HDTV) to CIF.

An example video frame of each clip is shown in Table 2, and Table 3 describes each audio-video clip.⁴ The ten clips were split into Pool A, which contained voice only (i.e., single person talking), and Pool B, which contained music and clips containing a mixture of sounds. Hockey1 contained crowd noise and an announcer talking; and music2 contained music and a person talking. Differential delay (related to “lip synchronization”) was not studied in this experiment.

Table 2. Example Video Frame from each Source Clip

| | | | | |
|---|---|---|--|---|
|  |  |  |  |  |
| cartalk1 | cchart2 | drmfeet | fish2 | guitar3 |
|  |  |  |  |  |
| hockey1 | music2 | presents3 | rfdev2 | stadpan |

⁴ Most of these source sequences are available free of charge for research purposes on the Consumer Digital Video Library (CDVL). The CDVL is available at www.cdvl.org.

Table 3. Audio-Video Sequence Descriptions and Pools

| Sequence | Pool | Video Recording | Audio Recording | Audio Type |
|-----------|------|--|---|------------|
| cartalk1 | A | Sony Z1U, HDTV 1080i, 29.97fps | Sony Z1U external mic | voice only |
| cchart2 | A | Sony DXC-MX7 3ccd, recorded on D-5 in 525-line/NTIA format | external mic on table below speaker | voice only |
| drmfeet | B | Sony Z1U, HDTV 1080i, 29.97fps | Sony Z1U external mic | music only |
| fish2 | A | Sony Z1U, HDTV 1080i, 29.97fps | Harvard balanced sentences recorded in sound isolation room | voice only |
| guitar3 | B | Sony Z1U, HDTV 1080i, 29.97fps | Sony Z1U external mic | music only |
| hockey1 | B | VHS recording | VHS recording with crowd noise | mixed |
| music2 | B | Sony Z1U, HDTV 1080i, 29.97fps | Sony Z1U external mic | mixed |
| presents3 | B | VHS recording | VHS recording | music only |
| rfdev2 | A | Sony DXC-MX7 3ccd, recorded on D-5 in 525-line/NTIA format | external mic on table below speaker | voice only |
| stadpan | A | Camera unknown 720p, 59.94 fps | Harvard balanced sentences recorded in sound isolation room | voice only |

Table 4 identifies the Hypothetical Reference Circuits (HRCs) used to encode the video, and Table 5 identifies the HRCs used to encode the audio. A Hypothetical Reference Circuit is a fixed combination of an encoder operating at a given bit-rate, network condition, and decoder. Video was encoded using the H.263, H.264 (also called MPEG-4 part 10), Windows Media version 9 (also called VC-1), and MPEG-2. Audio was encoded using MPEG-3 Audio, PCM, and Windows Media Audio (WMA). The range of audio impairments and video impairments

were chosen separately. The range of bit rates (and thus quality) was intended to represent a wide range of impairments that might realistically be encountered in systems deployed today.

This experiment simultaneously examined two full matrices of scenes, audio HRCs, and video HRCs. The audio-video clips in Pool A were paired with video HRCs V0, V1, V2, and V3, and also audio HRCs A0, A1, A2, and A3. The audio-video clips in Pool B were paired with video HRCs V0, V4, V5, and V6, and also audio HRCs A0, A4, A5, and A6. Thus, subjects were shown sixteen versions of clip cartalk1 (i.e., every combination of four audio HRCs and four video HRCs).

Table 4. Video HRCs

| Video Coder | Parameters | Condition Name |
|--------------------|-------------------------|-----------------------|
| Original Video | Uncompressed | V0 |
| H.263 | 600 kbps video bit rate | V1 |
| WMV9 / VC-1 | 125 kbps video bit rate | V2 |
| H.264 | 75 kbps video bit rate | V3 |
| MPEG-2 | 800 kbps video bit rate | V4 |
| H.264 | 250 kbps video bit rate | V5 |
| WMV9 / VC-1 | 75 kbps video bit rate | V6 |

Table 5. Audio HRCs

| Audio Coder | Parameters | Condition Name |
|--------------------|------------------------|-----------------------|
| Original Audio | Uncompressed | A0 |
| MPEG-3 | 24 kbps audio bit rate | A1 |
| BV16 | 16 kbps audio bit rate | A2 |
| WMA | 4 kbps audio bit rate | A3 |
| MPEG-3 | 32 kbps audio bit rate | A4 |
| BV16 | 16 kbps audio bit rate | A5 |
| WMA | 8 kbps audio bit rate | A6 |

2.2 Subjective Test Method

The subjective test was performed using the single stimulus Absolute Category Rating (ACR) method as described in [13], where viewers rated each sequence on a rating scale of: excellent, good, fair, poor, and bad. These items are typically mapped to the numbers 5, 4, 3, 2, and 1 respectively. Each subject viewed a unique randomization of the test sequences using a computer monitor and speakers. Subjects were allowed to take the test at their own pace and given a short break halfway through. For all subjects, the test took less than 45 minutes.

The viewing-listening environment was a sound-isolated chamber that conforms to ITU-T Rec. P.910. The subjects were allowed to view the sequences at a comfortable distance of their choosing. This was nominally 6-8 picture heights and was deemed typical of a desktop personal computer (PC) environment. The video was shown on an LCD monitor⁵ in CIF format. The audio was played on two speakers⁶ placed on either side of the LCD monitor, and pushed back slightly (i.e., the speakers were visible on either side of the monitor). The instructions read to the viewer are given in Appendix A, and the software used to administer the test is described in Appendix B. An automated program was used to ask the subject to rate the quality of each sequence. This automated program created a unique random ordering of clips for each subject.

Subjective ratings were gathered for two groups of subjective viewers (“Group X” and “Group Y”). Group X consisted of 32 naïve viewers, and Group Y consisted of 25 naïve viewers. Subjects from Group X were asked to rate the overall “multimedia” quality. Subjects from Group Y were asked to rate the same audio and video separately. That is, Group Y was given two sessions: one with video only and one with audio only. They were asked to rate the quality with the same rating scale as used in the audiovisual test. The order of these two sessions was randomized (i.e., subjects were randomly assigned to having either the video only session and the audio only session second, or vice versa).

⁵ This was a high quality LCD monitor.

⁶ Fostex 6301B speakers were used.

3 DATA ANALYSIS

3.1 Audio and Video Rated Together (Audiovisual Ratings)

This section describes the data analysis method for Group X, where the subjective viewers were asked to rate audio and video quality together (i.e., audiovisual rating).

When audio was at the highest quality possible (i.e., audio HRC A0), the mean audiovisual MOS scores, averaged per video HRC, had a spread of quality from 1.3 (between “bad” and “poor”) up to 4.5 (between “good” and “excellent”). This can be seen in Figure 5 and might be considered as being $Vq(Aq)$, though measured subjectively instead of objectively. When video was at the highest quality possible (i.e., video HRC V0), the mean audiovisual MOS scores, averaged per audio HRC, had a relatively narrow spread, ranging between 2.9 and 4.5. This can be seen in Figure 6 and might be considered $Aq(Vq)$. The narrow range of audio HRC MOSs (2.9 to 4.5) along with the wider range of video HRC MOSs (1.3 to 4.5) demonstrates that the video quality was the prevailing influence on the audiovisual quality score.

Put another way, the audio impairments and the video impairments spanned different ranges of quality. This imbalance became obvious to the experiment designers after the fact (i.e., when viewing and listening to the clips after examination of the subjective data), and resulted from separately choosing the audio and video impairments. All further conclusions from this experiment must take this disparity into account.

The average MOSs were computed by averaging the audiovisual MOSs of all clips for a given HRC. For example, the average MOS for HRC condition V0 was computed by averaging all the audiovisual MOSs of all clips which had a video HRC of V0. Likewise, the average audiovisual MOS for HRC condition A0 was computed by averaging all the MOSs of all clips which had an audio HRC of A0.

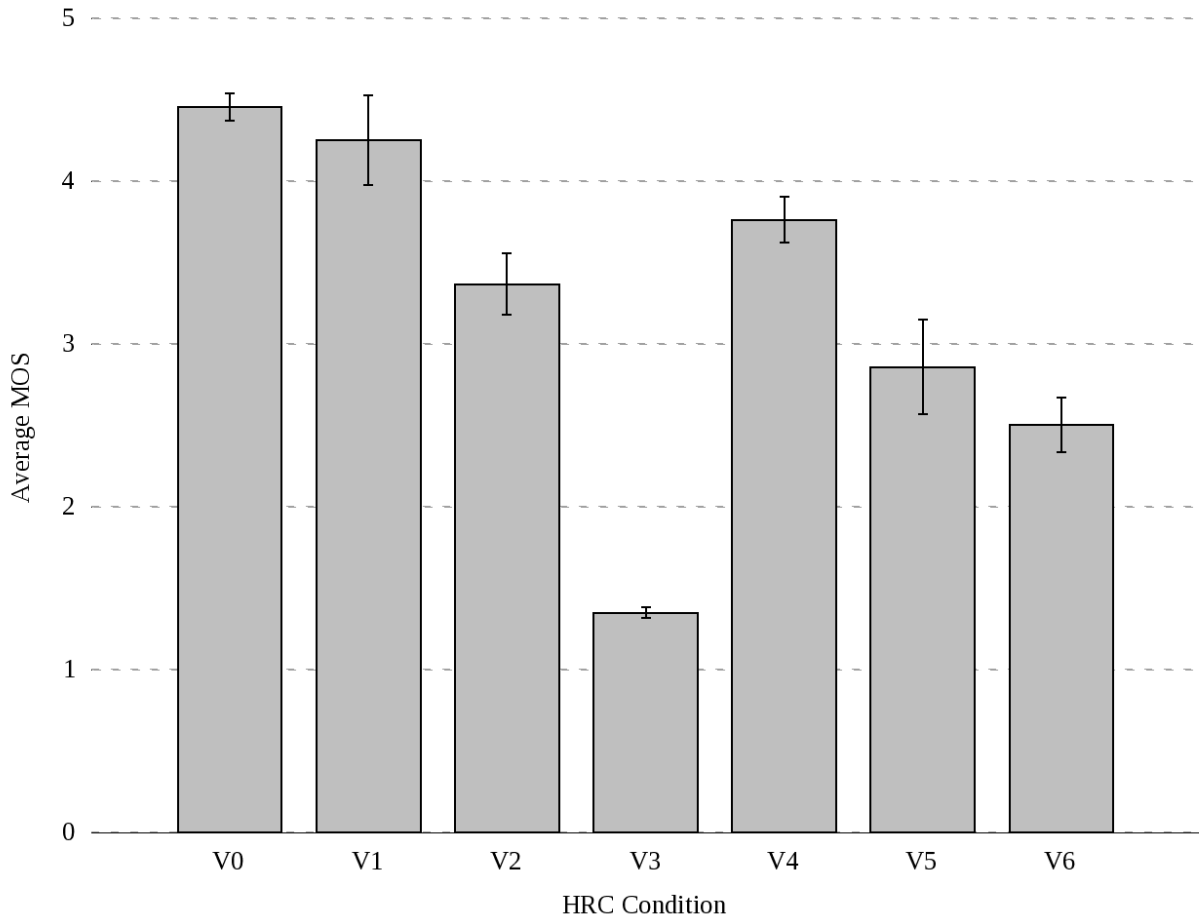


Figure 5. Mean audiovisual MOS when audio was at highest quality, averaged for each video HRC. This indicates the quality of each video HRC, influenced by the original audio quality.

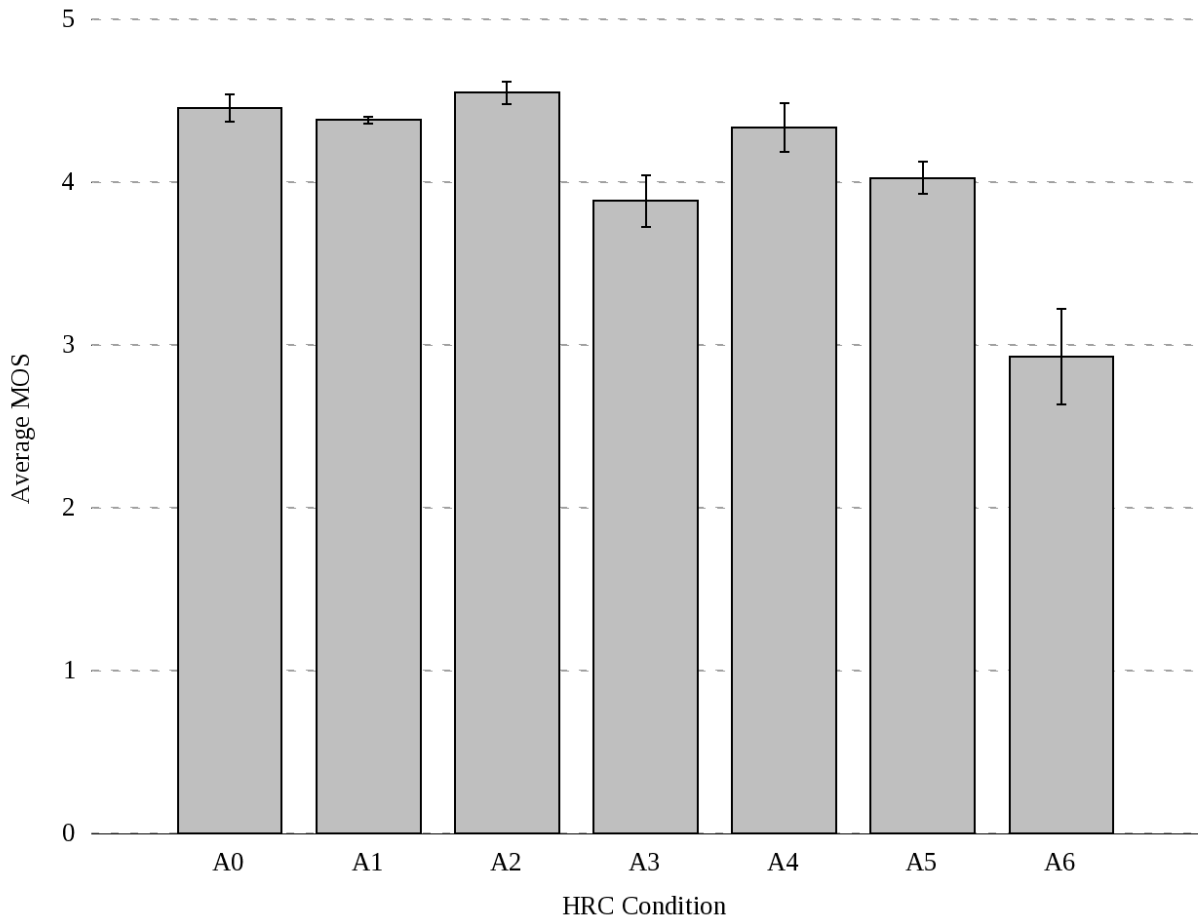


Figure 6. Mean audiovisual MOS when video was at highest quality, averaged for each audio HRC. This indicates the quality of each audio HRC, influenced by the original video quality.

This experiment consisted of two full matrices consisting of five scenes, each with four video HRCs and four audio HRCs. This design lends itself well to Analysis of Variance (ANOVA), which can be used to separate the impact of video and audio from the overall quality score. ANOVA indicated the following distribution of the variance of subjective data:

- 78% from differences between video HRCs
- 10% from differences between audio HRCs
- 4% from differences between individual clips
- 4% from interactions between the video HRCs and the individual clips
- 2% from interactions between the audio HRCs and the individual clips
- 1% from interactions between the video HRCs and audio HRCs

If we consider only the sequences containing speech, the response of subjective scores to conditions becomes even more one sided.

- 90% from differences between video HRCs
- 3% from differences between audio HRCs
- 2% from differences between individual clips

Note that for audio coding, the “single person talking” case is quite easy to code. This very simple case resulted in most audio “impairments” being nearly transparent and nearly identical. The lowest quality audio HRC (A3) was slightly muffled, but this did not seem to impact quality (that is, no one cared). Thus, all “voice only” impairments were nearly identical.

The music and mixed-audio cases were typically (but not always) more challenging for the audio codecs, and thus yielded more interesting results. Most of A4 and A5 impairments were nearly transparent. A6 produced impairments that included the loss of high pitches and/or warbling audio. This impairment was objectionable for some sequences (guitar3, hockey1, and music2) but not others (drmfeet and presents3).

Table 6 shows the two main effects of the ANOVA for each individual clip. The column “Notes” provides some further details regarding that sequence.

Table 6. ANOVA of individual audio-video clips

| Sequence | Audio Type | ANOVA | Notes |
|-----------|------------|----------------------------------|---|
| cartalk1 | voice only | Video HRC: 95% Audio HRC: 4% | Dramatic rendition of boy's complaint regarding sister. |
| cchart2 | voice only | Video HRC: 90% Audio HRC: 8% | Single person talking. |
| drmfeet | music only | Video HRC: 94% Audio HRC: 5% | Drums and cymbals. High pitches lost in A6, but minimal impact on music. |
| fish2 | voice only | Video HRC: 96% Audio HRC: 3% | Harvard balanced sentences. |
| guitar3 | music only | Video HRC: 45% Audio HRC: 48% | Soft guitar music, picked. High pitches lost in A6 impacted quality. |
| hockey1 | mixed | Video HRC: 43% Audio HRC: 51% | Crowd noise with announcer. A6 audio warbled. |
| music2 | mixed | Video HRC: 39% Audio HRC: 53% | Guitar, mandolin and banjo with talking. A6 audio warbled. |
| presents3 | music only | Video HRC: 89% Audio HRC: 1% | Modern "digital" music. High pitches lost in A6, but minimal impact on music. |
| rfdev2 | voice only | Video HRC: 95% Audio HRC: 4% | Single person talking. |
| stadpan | voice only | Video HRC: 100% Audio HRC: 0% | Video was very difficult to code. Harvard balanced sentences. |

3.2 Audio and Video Rated Separately

This section describes the data analysis method for Group Y, where the subjective viewers were asked to rate audio and video quality separately (i.e., video rating and audio rating). Each viewer/listener was presented stimuli of video only and audio only in separate sessions. Note that this was the same audio and video that was presented together to Group X.

A linear regression analysis was performed on the ratings of Group Y against the ratings of Group X. The purpose of this analysis was to examine how the separate ratings of audio quality and video quality given by Group Y might predict the audiovisual quality ratings given by Group X. This analysis shows that audiovisual quality ratings can be predicted from just video quality ratings with a correlation coefficient (ρ) of 0.92; Figure 7 shows the regression and its errors.

When the audio ratings alone were used to predict the audiovisual rating, ρ was 0.34; Figure 8 shows the regression and its errors. Using a cross term multiplying the audio and video quality ratings, the audiovisual quality ratings can be predicted with $\rho = 0.93$; Figure 9 shows the regression and its errors. Furthermore, when three terms are used (video quality ratings, audio quality ratings, and the audio-video cross term), ρ increases to 0.97; Figure 10 shows the regression and its errors. The linear combination equations and correlation coefficients for each case are presented in Table 7.

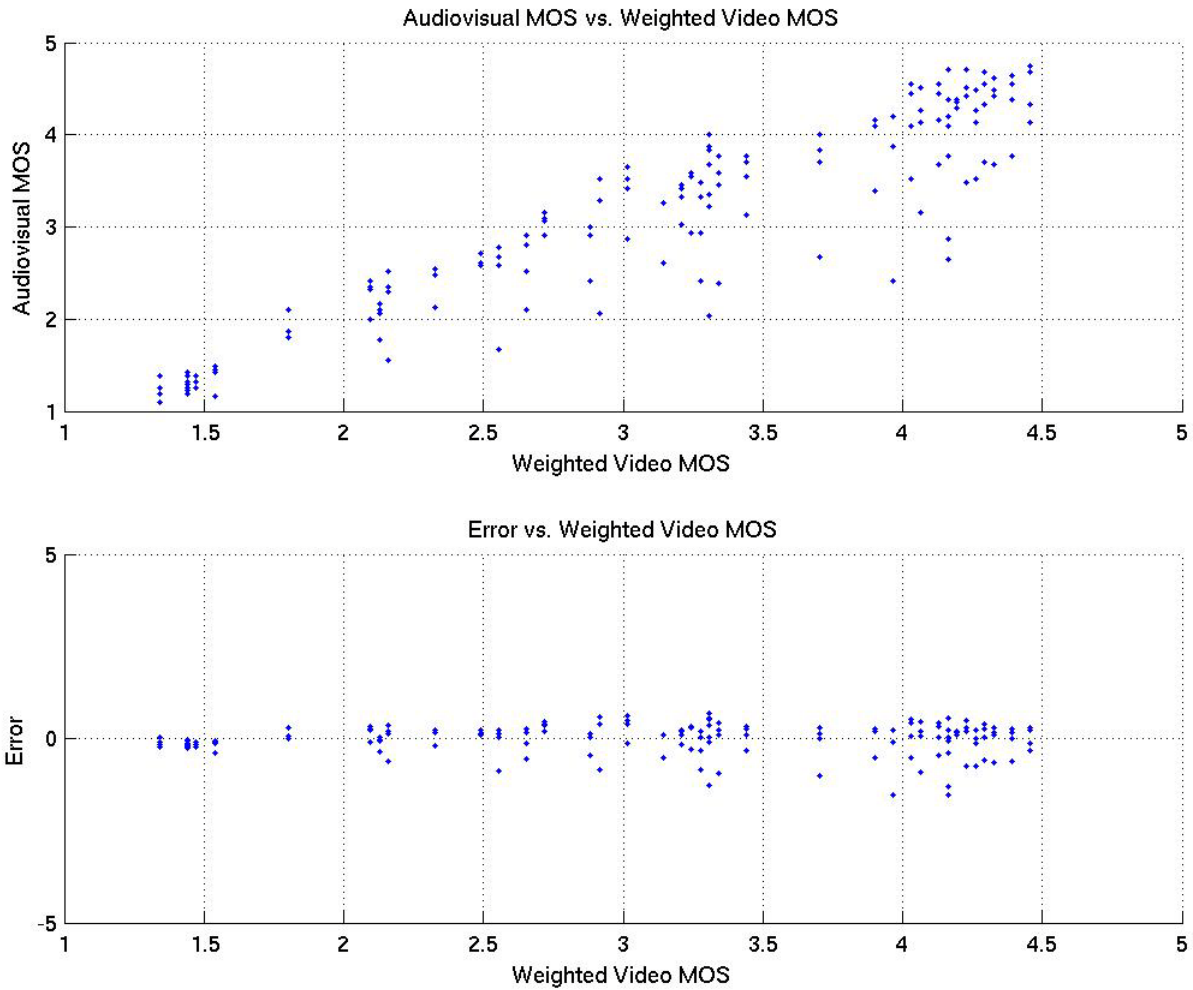


Figure 7. Audiovisual MOS predicted by weighted video MOS (top), and errors (bottom).

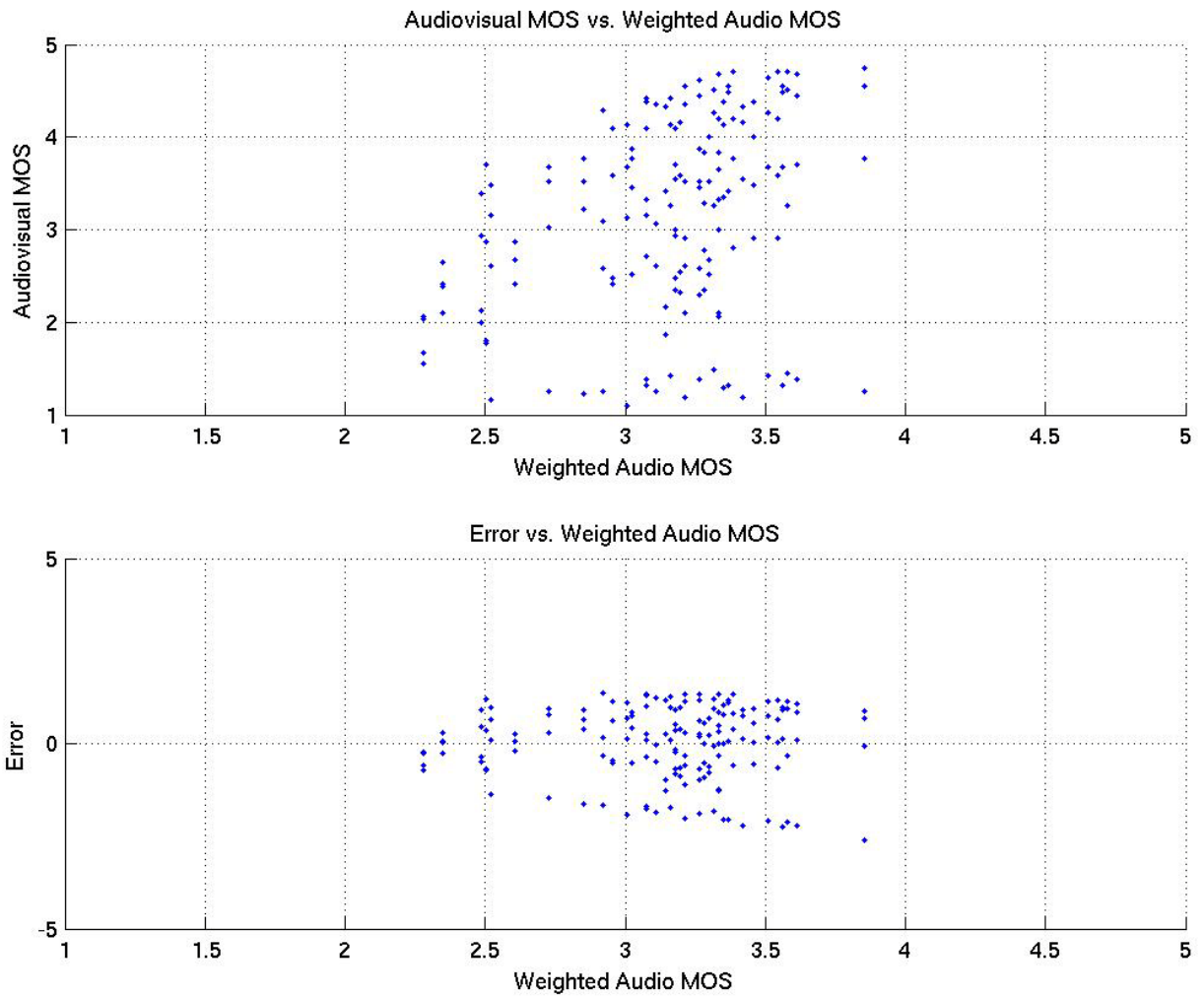


Figure 8. Audiovisual MOS predicted by weighted audio MOS (top), and errors (bottom).

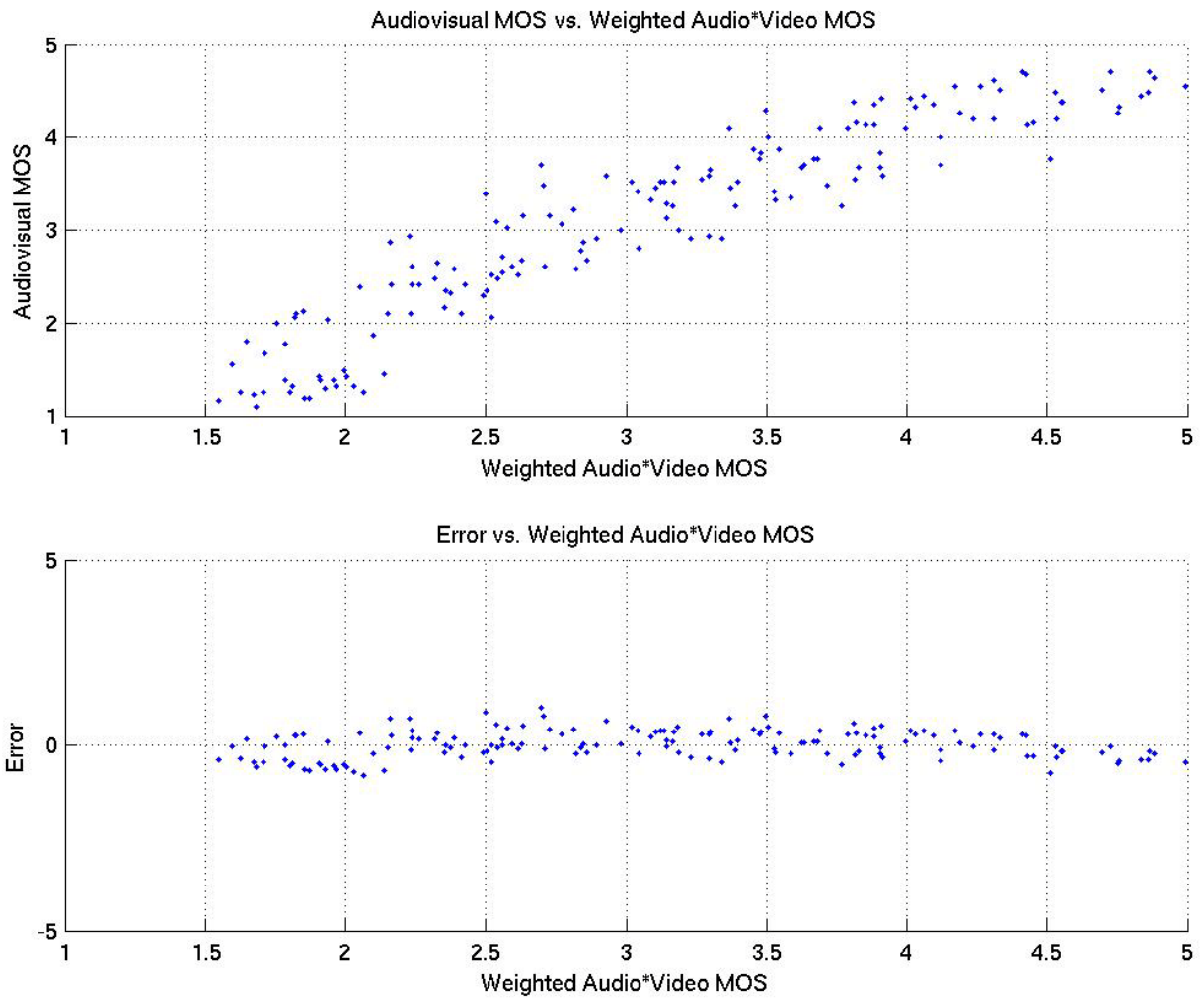


Figure 9. Audiovisual MOS predicted by weighted audio \times video MOS (top), and errors (bottom).

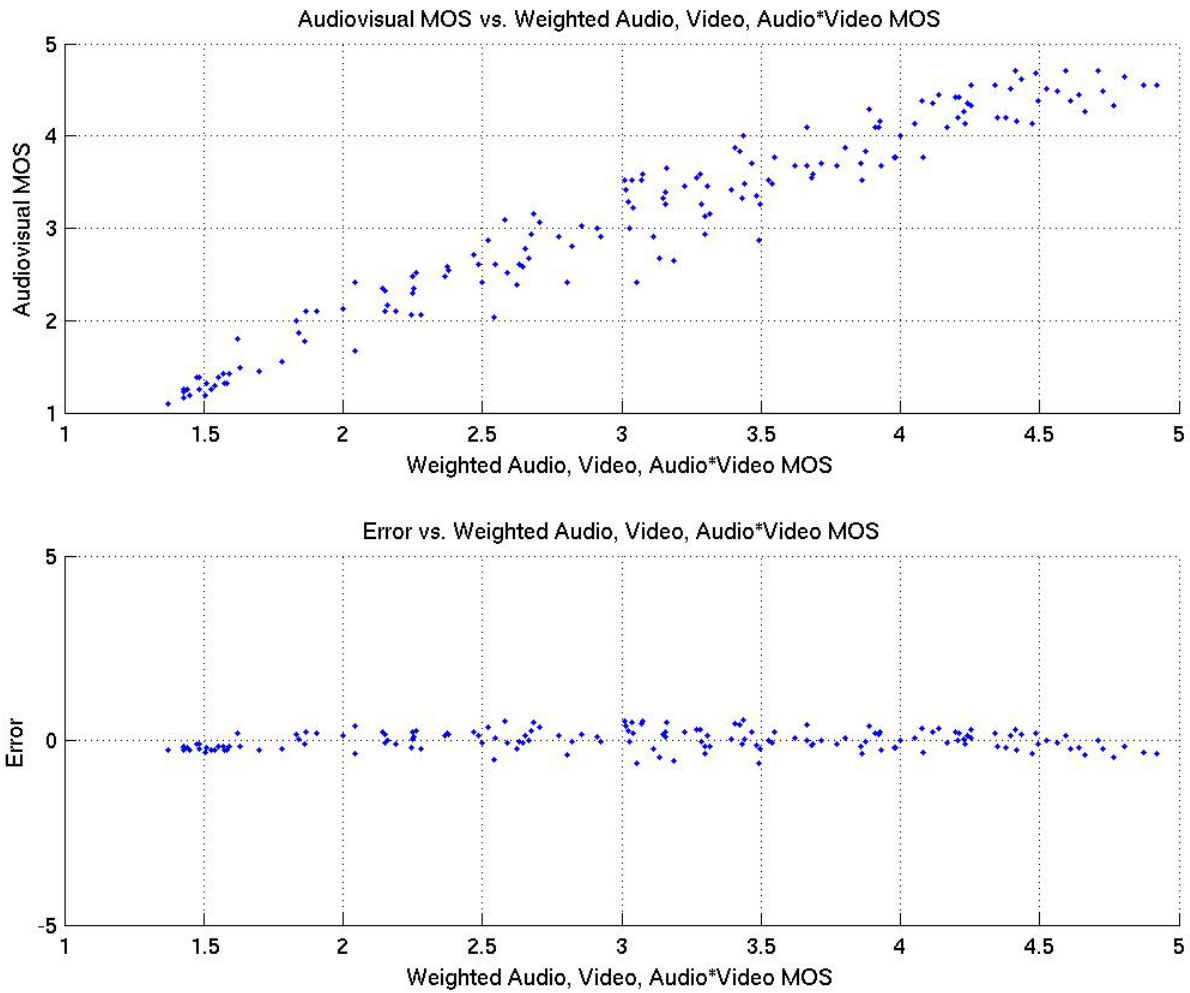


Figure 10. Audiovisual MOS predicted by weighted audio, video, and audio \times video MOS (top), and errors (bottom).

Table 7. Linear combination equations, ρ , and ρ^2 of experiment.

| Model No. | Model | ρ | ρ^2 | Terms |
|-----------|--|--------|----------|---------------------------------------|
| 1 | $\hat{s}_{av} = 0.5209 + 0.8201s_v$ | 0.92 | 0.85 | Video only |
| 2 | $\hat{s}_{av} = 1.7407 + 0.4332s_a$ | 0.34 | 0.12 | Audio only |
| 3 | $\hat{s}_{av} = 1.1096 + 0.1959(s_v \times s_a)$ | 0.93 | 0.86 | Audio \times Video |
| 4 | $\hat{s}_{av} = 0.7500 - 0.0452s_a + 0.3882s_v + 0.1250(s_v \times s_a)$ | 0.97 | 0.94 | Audio, Video and Audio \times Video |
| 5 | $\hat{s}_{av} = -0.5875 + 0.3599s_a + 0.8037s_v$ | 0.92 | 0.85 | Audio & Video, separately |

4 CONCLUSIONS

The full matrix of audio versus video was very useful, as it allowed for ANOVA to be performed effectively. Video HRCs evenly spanned a wide range of impairments. The results of this study show that the overall MOS for this set of audiovisual clips can be predicted rather well from the video MOS alone. Since the range of impairments for the audio HRCs was not as large as the video HRC range of impairments, it is unclear whether this indicates underlying truth, or was merely a consequence of the unbalanced test design. The choice of audio (which is 50% voice only) may have had a large influence as well.

While the linear regression analysis of Group Y's MOS shows that for this particular experiment, audiovisual quality can be predicted by the video MOS alone, it was most accurately predicted by a linear combination of the video MOS, audio MOS, and audio-video MOS cross term for this particular experiment.

The differences in the models from the experiments discussed in [5], [10] –[12] and this experiment (see Table 1) may be caused by differences in the source material, the impairments types, the subjective testing procedures, the task examined, or relative ranges of audio quality and video quality examined. More work is needed in this area.

5 FUTURE WORK

For future testing, we plan to minimize the scenes that contain voice only in favor of scenes with a variety of different types of music, background noise, multiple talkers, and mixed content. These results indicate that the type of music may impact the codec's performance, so a wide variety of instruments is desirable. We plan to conduct follow-on experiments in which voice-only and music are included in the same session. The split in the current test made it difficult to figure out what effects were from the HRC choice and which were from the audio type. Finally, the range of audio and video impairments should be chosen by watching and listening to the impaired audio-video sequences, rather than trying to make decisions on each aspect separately. This will help allow the ranges of audio and video MOSs be more comparable to each other.

The next experiment in the series of multimedia quality studies will be in the area of HDTV. The first HDTV study will focus on the effects of the perceived audio and video quality on the combined perceived multimedia quality, with no differential delay.

6 REFERENCES

- [1] K. Brunnström, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics," *IEEE Signal Processing Magazine*, vol. 97, May 2009.
- [2] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," Geneva, Switzerland, 2003 (available: www.itu.org).
- [3] M. Frater, J. Arnold, and A. Vahedian, "Impact of audio on subjective assessment of video quality in videoconferencing applications," *IEEE Transaction on Circuits and Systems for Video Technology*, September 2001.
- [4] M. Hollier et al, "Multi-modal perception," *BT Technology Journal*, January 1999.
- [5] C. Jones, D. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," Record of the 6th IEEE International Workshop on Quality of Service, Napa, CA, May 18-20, 1998.
- [6] K. Nakazono, "Frame rate as a QoS parameter and its influence on speech perception," *Multimedia Systems*, 1998.
- [7] "Requirements for an objective perceptual multimedia quality model," ITU-T Recommendation J.148, 2003 (available: www.itu.org).
- [8] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation P.862, 2001.
- [9] "Relative timing of sound and vision for broadcasting," ITU-R Recommendation BT.1359-1, 1998.
- [10] "Relations between audio, video, and audiovisual quality," ITU-T Contribution COM 12-19-E, February 1998. Contributed by KPN Research, Netherlands.
- [11] "Combined A/V model with Multiple Audio and Video Impairments," ANSI-Accredited Committee T1 Contribution, T1A1.5/94-124, April 10, 1995. Contributed by Bellcore, USA.
- [12] "Multimedia opinion model based on media interaction of audio-visual communications," ITU-T Contribution COM 12-D36-E, January 2005, University of Tsukuba, Japan.
- [13] "Subjective video quality assessment methods for multimedia applications," ITU-T Recommendation P.910, 1999.

APPENDIX A: VIEWER INSTRUCTIONS

The following are the viewer instructions for members of Group X, which rated the overall audiovisual quality. Members of Group Y were given similar instructions for the video only and audio only sessions.

Thank you for coming in to participate in our study. This study is about the quality of audiovisual multimedia and will assist us in the evaluation of multimedia transmission systems.

In this experiment, you will be presented with short audiovisual sequences. Each time a sequence is presented, judge its quality using the mouse of the computer and selecting one of the five levels on the following scale.

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

Your evaluation must reflect your opinion of the overall combined audiovisual quality.

We will start with five practice sequences together. After that, the experiment will consist of two blocks of sequences that will each take approximately 20 minutes to complete. You will be notified when you complete the first block, at which point you should take a break. You may leave the test chamber, [and I will have some snacks and beverages for you].

Observe and listen carefully to the entire clip before making your judgment. Keep in mind that you are rating the **quality** of each clip, **not its content**. There is no right or wrong answer to this; we are interested in your opinion.

Do you have any questions before we begin?

[perform practice session together and answer any questions]

APPENDIX B: TEST SOFTWARE DESCRIPTION

The multimedia testing software uses the same Java™ graphical user interface (GUI) for both the administrator and the subjects. The interface allows the administrator to customize the test. The administrator has control over options such as the desired video output drivers (DirectX or OpenGL), the ability to run the test on one or two monitors, the quality scale (five or nine level scales are available), and whether the subject is rating audio files, video files, or both. These settings can be saved so that later tests can be run with those identical settings. Figure B-1 shows a screen shot of the interface where the administrator specifies these variables.

The desired video and/or audio clips are loaded through the GUI by the administrator. Clips can be loaded into either the practice space (allowing users to get a feel for the testing procedure without the results being counted) or into the testing space (in which the user's ratings are registered and saved to a numbered file on the hard drive). The screen used to choose the audio-video files is shown in Figure B-2.

Once the test environment is created, pressing "Start Test" in the interface shown in Figure B-1 starts the subjective test interface. An introductory screen is presented while the viewer is seated. When the subject is ready to begin, an on-screen button is pressed to play the practice clips. The video and audio files are played using the freeware player "MPlayer," using command line calls from within the Java GUI. A different player can be used, provided that it has a command line interface and suitable GUI. After viewing and/or listening to the sequence, the subjects choose a rating based on what they saw and/or heard (see Figure B-3). After the practice session, the software pauses to allow questions to be asked. Then, the subjects are presented with the audio-video sequences from the experiment.

The subject's opinion scores are saved to a file and associated with that subject's identification number. Subject identification numbers are not associated with subjects' names due to privacy concerns.

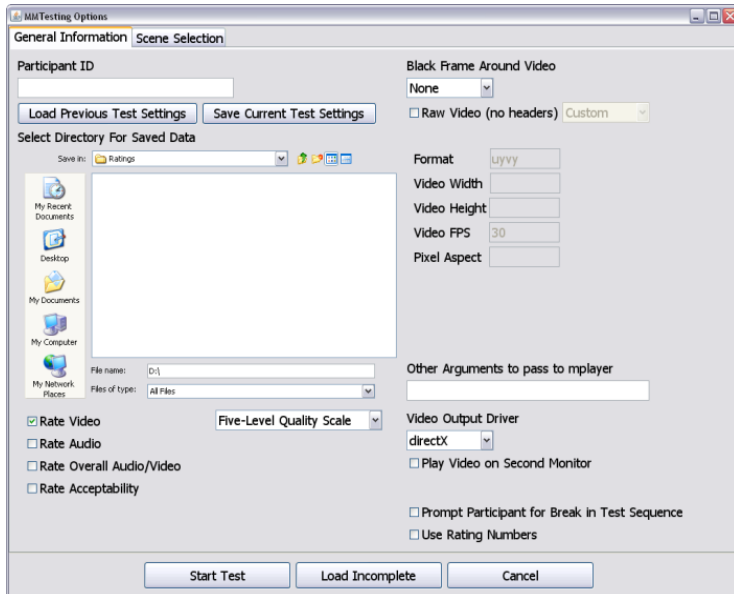


Figure B-1. Subjective test control program interface used to specify type of experiment.

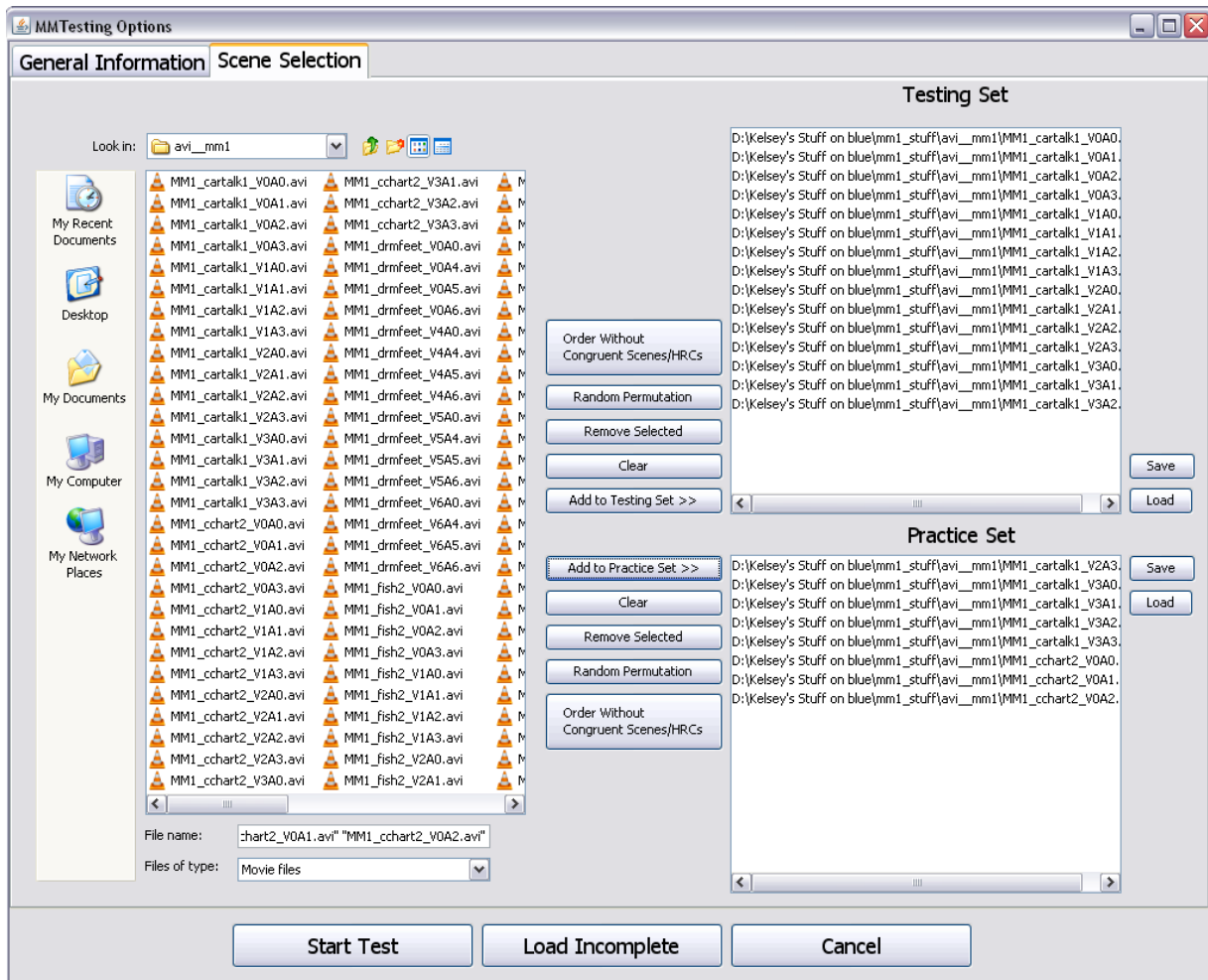


Figure B-2. Subjective test control program interface used to specify test sequences and practice session.



Figure B-3. The rating screens for Group X (left) and Group Y (right) showing the MOS scales.

FORM NTIA-29
(4-80)

U.S. DEPARTMENT OF COMMERCE
NAT'L. TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

| | | |
|---|---|--|
| 1. PUBLICATION NO. TR-10-472 | 2. Government Accession No. | 3. Recipient's Accession No. |
| 4. TITLE AND SUBTITLE Relating Audio and Video Quality Using CIF Video | | 5. Publication Date September 2010 |
| | | 6. Performing Organization Code |
| 7. AUTHOR(S) Mark A. McFarland, Margaret Pinson, Carolyn Ford, Arthur Webster, William Ingram, Scott Haines, Kelsey Anderson | | 9. Project/Task/Work Unit No. 3139012-300 |
| 8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305 | | 10. Contract/Grant No. |
| | | 12. Type of Report and Period Covered |
| 11. Sponsoring Organization Name and Address National Telecommunications & Information Administration Herbert C. Hoover Building 14 th & Constitution Ave., NW Washington, DC 20230 | | |
| 14. SUPPLEMENTARY NOTES | | |
| 15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.) NTIA/ITS has conducted a series of studies to quantify the effects that individual audio and video qualities have on the overall Mean Opinion Score (MOS) for a given set of audiovisual clips. The experiment described in this report studies the effects which that synthesis of audio and video quality have on a subject's overall MOS. The results of this study show that the overall MOS for this set of audiovisual clips can be predicted rather well from the video MOS alone. This result will not necessarily be valid for other choices of audiovisual material. | | |
| 16. Key Words (Alphabetical order, separated by semicolons) audio quality; subjective testing; video quality | | |
| 17. AVAILABILITY STATEMENT <input checked="" type="checkbox"/> UNLIMITED. | 18. Security Class. (This report) Unclassified | 20. Number of pages 24 |
| | 19. Security Class. (This page) Unclassified | 21. Price: |

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities. Subsets of this series include:

NTIA RESTRICTED REPORT (RR)

Contributions that are limited in distribution because of national security classification or Departmental constraints.

NTIA CONTRACTOR REPORT (CR)

Information generated under an NTIA contract or grant, written by the contractor, and considered an important contribution to existing knowledge.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail info@its.blrdoc.gov.

This report is for sale by the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161, Tel. (800) 553-6847

