Margaret H. Pinson, William Ingram, and Arthur Webster

Audiovisual Quality Components

An analysis

© INGRAM PUBLISHING

The perceived quality of an audiovisual sequence is heavily influenced by both the quality of the audio and the quality of the video. The question then arises as to the relative importance of each factor and whether a regression model predicting audiovisual quality can be devised that is generally applicable.

INTRODUCTION

This article analyzes subjective experiments that explore the relationship between audio quality and video quality, measured separately, and the overall quality of an audiovisual experience. This topic has been explored by at least 12 previous experiments [1]–[11] and one additional experiment described here. Each of these experiments produced a model that mapped audio quality (***a***) and video quality (***v***) to the overall audiovisual quality (***av***). These models are all linear; however, the terms and the values of the coefficients used in the model differ from one experiment to the next.

The goal of this analysis is to describe a flexible audiovisual model that can be applied to a wide range of impairments, applications, source material, and video resolutions. Due to practical limitations, any one subjective experiment can only explore a limited portion of this larger problem. Thus, each experiment contributes to the learning experience: some insights into audiovisual quality, knowledge of what worked well with this experiment, and feedback as to what should be changed for future experiments.

This article contains two main sections. The first describes an experiment conducted at the Institute of Telecommunication Sciences (ITS) in 2010. The second summarizes this and previous experiments, then jointly analyzes the data from all of the summarized experiments to see what conclusions may be

Digital Object Identifier 10.1109/MSP.2011.942470
Date of publication: 1 November 2011

reached. These analyses focus on relative audiovisual qualities (e.g., the quality of one presentation compared to that of another). While important, context effects from specific applications, devices and environments are not considered here. Other studies address context effects (e.g., [17]).

THE GOAL OF THIS ANALYSIS IS TO DESCRIBE A FLEXIBLE AUDIOVISUAL MODEL THAT CAN BE APPLIED TO A WIDE RANGE OF IMPAIRMENTS, APPLICATIONS, SOURCE MATERIAL, AND VIDEO RESOLUTIONS.

ITS 2010 AUDIOVISUAL EXPERIMENT

MOTIVATIONS FOR EXPERIMENT DESIGN

The experiment described in this article was designed to replicate the experiment performed by ITS in 2009 [11] while addressing that experiment's two primary flaws and using high-definition TV (HDTV). The samples for both the ITS 2009 experiment and this, ITS 2010, consisted of a set of audiovisual sequences, where the audio and video were impaired separately. These separate impairments were then combined in all combinations, such that subjects were presented with a full matrix (i.e., all audio impairments combined with all video impairments). This full matrix allows for interesting analysis of variance (ANOVA) on the audiovisual data, to separate the relative impact of audio and video within the overall audiovisual quality scores.

Analysis of ITS 2009 (performed using common intermediate format (CIF) video, 352 × 288) indicated that the video impairments spanned a much wider range of quality than the audio impairments. The question thus arose as to whether the greater weight on video in the audiovisual model was unduly influenced by this unequal distribution. Thus, the ITS 2010 study examined the hypothesis that, if the audio quality spanned nearly the same range as the video quality, then the audio and video quality would be equally important in the overall audiovisual quality.

The ITS 1998 experiment [4] took a different approach. It chose audio impairments that naturally and logically matched with video impairments. This constrained the range of audio and video impairments to those seen in common usage. Since audio requires much lower bitrates than video for CIF resolution and above, the audio quality impairments likewise spanned a limited range of quality. Therefore, audio and video impairments for the ITS 2010 experiment were selected to span approximately the same range of quality.

Another flaw seen in ITS 2009 concerned the types of audio samples used. That experiment intentionally contained 50% sequences with audio consisting of a single person talking. This choice was made out of respect for the extensive research efforts previously conducted using objective models that measure the quality of audio containing a single person talking with no background noise. From an audio compression standpoint, a single person talking is extremely easy to code. Thus, relatively little new information was learned when making comparisons within these audiovisual sequences.

Of the ten audiovisual sequences examined by ITS 2009, only three were associated with an interesting balance such that both the audio quality and the video quality significantly impacted the overall audiovisual quality. For the other seven sequences, the video quality dominated (i.e., video quality explained 89–100% of the distribution of the variance in the subjective data). The three more interesting audio samples contained soft guitar music, crowd noise with an announcer, and music with some talking. In contrast, the ITS 2010 experiment was designed to contain audio that minimizes single person talking samples, and instead emphasizes more complicated audio (e.g., a single person talking with music in the background). A variety of different music types were included, in the hopes that different instruments might elicit different weightings of audio and video quality, thus better representing a wide range of all types of audio.

The decision to use HDTV was motivated by the increasing importance of higher-resolution video. As previous ITS experiments had examined only lower resolution video, examining HDTV would add value.

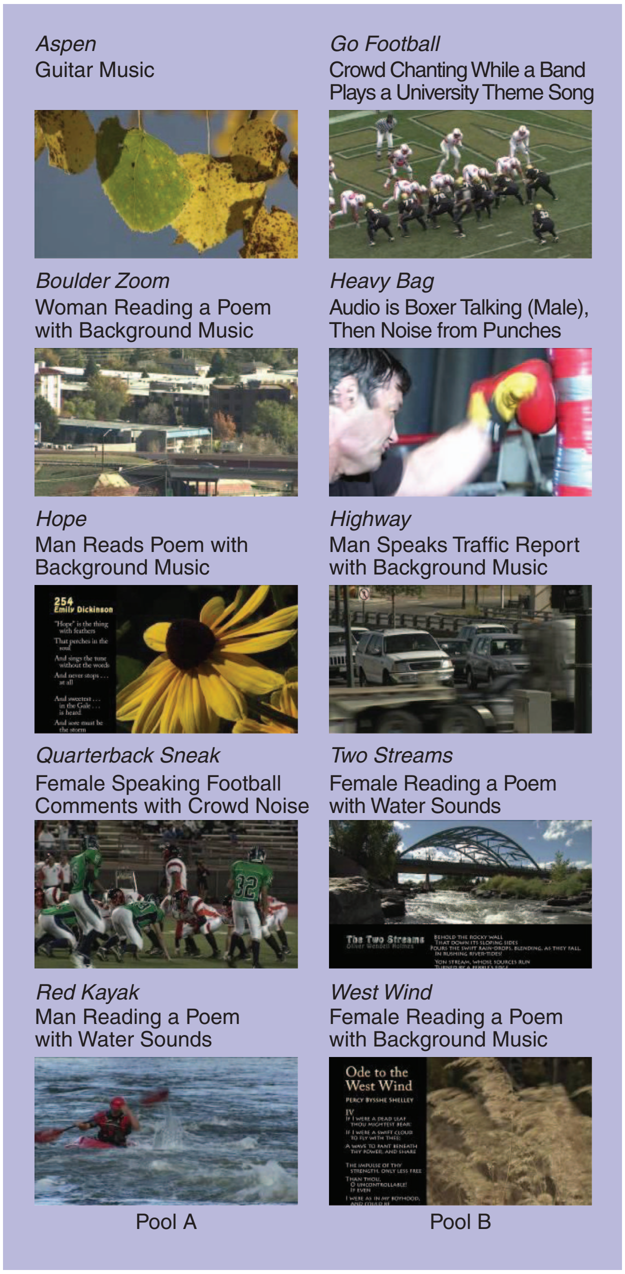
AUDIOVISUAL SEQUENCES

This experiment contained ten audiovisual sequences of 15 s each. Each original sequence contained video that, when visually inspected by an expert, appeared to be of good or better quality. These video sequences were carefully chosen to span a wide range of coding difficulty and a variety of visual characteristics (e.g., scrolling text, fast motion, rapid scene cuts, and random motion). The audio associated with the video sequences was, in most cases, dubbed after the video was created. For most of the sequences, the audio consisted of a single person speaking in a sound-isolated room, combined with either music or background noise. The audio related meaningfully to the video (e.g., a poem *Ode to the West Wind* was paired with grass blowing in the wind). The music represented a wide range of styles and instruments. The sequences were divided into two pools of five, which contained approximately equivalent video and audio characteristics.

Figure 1 shows a sample frame from each sequence, along with a description of the audio and the division of sequences into Pool A and Pool B. These videos were filmed in either 1080p30 or 1080i60 (that is, 1920 × 1080 in either progressive 29.97 frames per second (fps) or interlaced 59.94 fields per second). The audio was recorded in either stereo or mono, at 48 kHz. These audiovisual sequences can be viewed in the Consumer Digital Video Library (www.cdvl.org). They are available for research purposes from that site.

HYPOTHETICAL REFERENCE CIRCUITS

One of the design goals was to produce audio and video hypothetical reference circuits (HRCs) such that each set of


Aspen
Guitar Music
Go Football
Crowd Chanting While a Band Plays a University Theme Song
Boulder Zoom
Woman Reading a Poem with Background Music
Heavy Bag
Audio is Boxer Talking (Male), Then Noise from Punches
Hope
Man Reads Poem with Background Music
254 Emily Dickinson
Highway
Man Speaks Traffic Report with Background Music
Quarterback Sneak
Female Speaking Football Comments with Crowd Noise
Two Streams
Female Reading a Poem with Water Sounds
The Two Streams
Red Kayak
Man Reading a Poem with Water Sounds
West Wind
Female Reading a Poem with Background Music
Ode to the West Wind
Percy Bysshe Shelley
Pool A
Pool B

[FIG1] Sample frame from each video sequence displayed below a brief description of the audio.

impairments spanned the full range of quality, from excellent to bad. This experiment included the original video plus six video impairments, and the original audio plus three audio impairments. The audio impairments were Advanced Audio Coding (AAC) at 16, 32, and 48 kb/s. The video impairments were Advanced Video Coding (AVC) at 2, 3.5, and 6 Mb/s, and MPEG-2 at 6, 8.5, and 12 Mb/s. (AAC is also known as ITU-T H.264 [12] and ISO/IEC 14496-10 [13].) The video impairments were chosen from those used in [14], as this provided precise feedback on the expected mean opinion scores (MOS). Reference [14] found that AVC at 3.5 Mb/s was statistically equivalent to MPEG 2 at 8.5 Mb/s. Both received an average MOS between fair and good on the Absolute Category Rating (ACR) scale; see [14, Fig. 3].

Audio impairments were chosen to approximately match the video impairments' range of quality, at the expense of realism. Note that the audio bit rates commonly associated with these video HRCs is high enough that even the lowest associated audio bit rate would cause little impairment to the audio. The selection of the audio and video impairments to span the range of quality was made by several video quality measurement engineers and based on subjective test results from previous experiments (e.g. [16]).

The three AVC encodings were paired with the five video sequences in Pool A (Aspen, Boulder Zoom, Hope, Quarterback Sneak, and Red Kayak). The three MPEG-2 encodings were paired with the five video sequences from Pool B (Go Football, Highway, Heavy Bag, Two Streams, and West Wind). The same three AAC audio impairments were paired with all sequences in both Pool A and Pool B. The pool designations serve only to keep track of which scenes were compressed with AVC and which with MPEG-2.

For each pool, a full matrix of audiovisual impairments was created: three audio impairments plus the original audio, by three video impairments plus the original video, for a total of 16 processed sequences. Each audio impairment was also included in isolation (i.e., audio only, four versions for each of ten sequences) and each video impairment was included in isolation (i.e., video only, four versions for each of ten sequences). Thus, the entire experiment included 240 samples: 160 audiovisual samples, 40 audio only, and 40 video only.

PRESENTATION TO SUBJECTS

Each subject observed and rated half of these sequences, drawn randomly from both Pool A and Pool B, using the ACR scale. Sequences were presented in a random order. The sessions were recorded to Blu-Ray discs. Three randomized splits of the data were created, such that each split included half of the original video sequences (i.e., half of the audiovisual sequences with original audio and original video; half of the original audio only sequences, and half of the original video only sequences). Each split was further broken into three subsessions A, B, and C each containing a random third of the sequences on the disc.

Each Blu-Ray disc contained the same training session, and therefore each subject saw the same training session prior to the test. Each subject saw the three subsessions specific to that disc (A, B, and C) in a random order (e.g., ABC, CAB, BCA), such that no more than three subjects were shown any particular order. Thus, in total there were six Blu-Ray discs each with subsessions A, B, and C specific to that disc:

- Split 1, Even
- Split 1, Odd
- Split 2, Even
- Split 2, Odd
- Split 3, Even
- Split 3, Odd.

Each pair of discs (even plus odd) contained all video sequences. The subjects watched the video on a TV-Logic LVM-460WD professional grade 46-in liquid crystal display (LCD) monitor in a sound-isolated room. They heard the audio over NHT Audio LLC speakers; main speaker model A-20 with subwoofer model B-20. A total of 54 naïve subjects ran through the experiment using the Blu-Ray discs. The subjects were primarily students from the local university who responded to an online advertisement. No subjects had participated in a similar test within the previous six months. Each subject's participation was limited to 90 min.

To analyze the reliability of this method, the following analysis was performed. For each of the 240 processed sequences, a comparison was made between the individual data resulting from one Blu-Ray disc, and the MOS for that clip computed from the other five Blu-Ray discs. The comparison was made using the Student's t-test at the 99% confidence level. For example, the Student's t-test computed with 99% confidence whether the individual opinion scores for Clip 1 on disc "Split 1, Even" appeared to be a random sample from a normal distribution with mean equal to the MOS for that clip from the other two discs where it appeared. This test was performed with the raw opinion scores as specified by the subjects (i.e., no scaling was performed).

Of the Student's t-test comparisons, 6% indicated a different mean at the 99% confidence level. This indicates a fair degree of reliability, given the known noise in subjective data and the impact of ordering effects. We expect some differences in the MOS from one viewer ordering to another, which is the motivation for maximizing the number of orderings presented to subjects. Thus, the viewers of all six Blu-Ray discs may reasonably be interpreted as one single set of viewers. (Note: when opinion scores from one disc are compared to another disc, this value rises to from 6% to approximately 10%.)

ITS 2010 ANALYSIS

Statistical analysis of this experiment will be presented in the next section.

SUMMARY OF ALL EXPERIMENTS

OVERVIEW

Due to practical limitations, any one subjective experiment can only explore a limited portion of the larger problem of describing a flexible model for describing audiovisual quality applicable to a wide range of sample types. Thus, each experiment provides a learning experience: some insights into audiovisual quality, knowledge of what worked well with this experiment, and feedback as to what should be changed for future experiments. This section will examine previous experiments as well as the ITS 2010 experiment and see what conclusions can be reached.

Table 1 summarizes the design of prior experiments that generated an equation. Experiments are listed in chronological order, because previous studies may have influenced later experiments. Each of these experiments performed separate subjective testing of the audio quality of audio-only sequences (i.e., no picture), the video quality of video-only sequences (i.e., silent), and the audiovisual quality of sequences containing some combinations of those audio and video samples. Most of the subjective experiments used a single stimulus methodology that measured people's opinions using MOS. While the majority of experiments mentioned in this article used single stimulus and MOS, one used double stimulus and differential MOS (DMOS). To avoid complicated language, the subjective data for that experiment will occasionally be incorrectly referred to as MOS. Where a single paper listed results from multiple experiments, each experiment is listed on a separate line of the table (e.g., the BT study).

Table 1 lists the video impairments, audio impairments, and the number of processed sequences that were used for the audiovisual portion of each experiment only. Additional impairments of video or audio sequences may have been used in the video-only or audio-only subexperiments and their analysis. The number of processed sequences listed includes the originals, if included in that experiment. Unless otherwise specified, the three types of stimuli (audio only, video only, and audiovisual) were presented in separate sessions. Taken together, these experiments span a wide range of applications, video resolutions, video impairments, audio impairments, and source material.

Table 2 summarizes the equations calculated from each experiment (column "Model") and the Pearson correlation of each model (column "ρ"). All models identified in the original papers that predict audiovisual quality using a combination of audio and video quality are shown. The column "Range of MOS" shows the relative range of MOS spanned by audio-only, video-only, and audiovisual subjective data. The column "Type Comparison" lists the Pearson correlation between audio-only MOS and audiovisual MOS as well as the Pearson correlation between video-only MOS and audiovisual MOS. The final column, "Dominant Factor," lists the subjective conclusion reached by the people conducting the experiment regarding the relative contributions of audio and video in the overall audiovisual quality. This information should be understood in the context of the experimental designs summarized in Table 1.

The models identified in Table 2 are numbered by the model type as follows:

$$1: \hat{y} = \alpha + \mu(\boldsymbol{a} \times \boldsymbol{v})$$
$$2: \hat{y} = \alpha + \beta\, \boldsymbol{a} \times \gamma \boldsymbol{v}$$
$$3: \hat{y} = \alpha + \gamma \boldsymbol{v} + \mu\, (\boldsymbol{a} \times \boldsymbol{v})$$
$$4: \hat{y} = \alpha + \beta\, \boldsymbol{a} + \gamma \boldsymbol{v} + \mu(\boldsymbol{a} \times \boldsymbol{v}).$$

The term $\boldsymbol{a}$ represents the measured audio quality MOS; $\boldsymbol{v}$ represents the measured video quality MOS; and $\boldsymbol{av}$ represents the measured audiovisual quality MOS. The term $\hat{y}$ represents the predicted audiovisual quality. Occasionally a paper will mention the accuracy of one of these model types but not the coefficients. In this case, the model coefficients are left Greek variables. France Telecom additionally presented a model that applied a logarithmic transformation to $\boldsymbol{av}$ and thus estimated

[TABLE 1] COMPARISON OF EXPERIMENT DESIGN FROM DIFFERENT LABORATORIES' INVESTIGATIONS INTO AUDIOVISUAL MODELS.

LABORATORY	FOCUS	DESIGN	SIZE	ENVIRONMENT	VIDEO	AUDIO	SCALE
BELLCORE 1993 [1]	ENTERTAINMENT TELEVISION NTSC	FULL MATRIX OF FIVE AUDIO ONLY BY FIVE VIDEO ONLY. ORIGINALS NOT RATED.	TWO ORIGINALS 50 PVS 18-S CLIPS	TELEVISION MONITOR (CRT) SPEAKERS	RANDOM NOISE, SIMULATED VIDEO CONTENT: ■ CONVERSATION WHILE WALKING ON A BUSY STREET (VARIETY AND MOTION) ■ CONVERSATION WHILE SITTING IN A LIBRARY (SOME HEAD-AND-SHOULDERS, LITTLE MOTION)	TEMPORALLY CORRELATED NOISE, SIMULATING A LOW BIT RATE VOICE CODER AUDIO CONTENT: SPEECH, MAY OR MAY NOT HAVE BACKGROUND NOISE (NOT SPECIFIED)	NINE-POINT SCALE, (EXCELLENT, GOOD, FAIR, POOR, UNSATISFACTORY) SIMILAR TO ACR
BELLCORE 1994 [2]	ENTERTAINMENT TELEVISION NTSC	IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]	SIMULATED BLURRING FROM A HORIZONTAL FIR LOW-PASS FILTER VIDEO CONTENT IDENTICAL TO BELLCORE 1993 [1]	MODULATED NOISE REFERENCE UNIT (MNRU) [15] AUDIO CONTENT IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]
BELLCORE 1995 [3]	ENTERTAINMENT TELEVISION NTSC	THREE FULL MATRICES, EACH THREE AUDIO ONLY BY THREE VIDEO ONLY	TWO ORIGINALS 54 PVS 18-S CLIPS	IDENTICAL TO BELLCORE 1993 [1]	VIDEO IMPAIRMENTS: ■ SAME AS [1] ■ SAME AS [2] ■ SIMULATED BLOCKINESS VIDEO CONTENT IDENTICAL TO BELLCORE 1993 [1]	AUDIO IMPAIRMENTS: ■ SAME AS [1] ■ SAME AS [2] ■ RANDOM NOISE AUDIO CONTENT IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]
ITS (1998) [4]	VIDEO-TELECONFERENCE (VTC) NTSC	AUDIOVISUAL IMPAIRMENTS CREATED, AND THEN SPLIT INTO AUDIO ONLY AND VIDEO ONLY AUDIOVISUAL BIT RATE 128–1,536 kb/S.	SIX ORIGINALS 48 PVS 5–9-S CLIPS	PC MONITOR (CRT) PC SPEAKERS	ANALOG, H.261, AND PROPRIETARY CODER VIDEO CONTENT: VTC (E.G., HEAD-AND-SHOULDERS, PEOPLE AT A TABLE, MAP WITH POINTER)	ANALOG, G.711, G.722, G.728, AND PROPRIETARY CODER AUDIO CONTENT: ■ SIX WITH SPEECH	ACR FIVE-POINT SCALE
FRANCE TELECOM/CNET 1998 [5]	VIDEO-TELECONFERENCE (VTC) PAL	FULL MATRIX OF FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	TWO ORIGINALS 32 PVS 10-S CLIPS	MONITOR AND LOUDSPEAKERS	ORIGINAL, CIF AND QCIF AT 12 OR 25 FPS FROM 172 TO 456 kb/S VIDEO CONTENT: VIDEOCONFERENCE	ORIGINAL, G.722,G.711, AND G.728 FROM 16 TO 56 kb/S AUDIO CONTENT: ■ TWO WITH SPEECH (ONE MALE, ONE FEMALE)	ACR FIVE-POINT SCALE
KPN RESEARCH 1997 [6], [7]	BROADCAST TELEVISION PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	TWO ORIGINALS 32 PVS 25-S CLIPS	TELEVISION MONITOR (CRT) STEREO LOUDSPEAKERS	SPATIAL FILTERING OF LUMINANCE SIGNAL IN HORIZONTAL DIRECTION. VIDEO CONTENT: ■ COMMERCIALS	CD QUALITY, BAND LIMITED TO WIDE BAND, AM RADIO, TELEPHONE QUALITY CONTENT NOT SPECIFIED.	ACR NINE-POINT SCALE
BT 2004 [8] EXPERIMENT #1	VIDEO-TELECONFERENCE PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	TWO ORIGINALS 32 PVS 5-S CLIPS ORIGINALS WERE RATED	TELEVISION MONITOR (CRT) SPEAKERS	EMULATED BLOCKINESS VIDEO CONTENT: LOW MOTION, LOW COMPLEXITY	MNRU LEVEL THREE TO 24 AUDIO CONTENT: ■ TWO WITH SPEECH (ONE MALE, ONE FEMALE)	DSCQS 100-POINT SCALE
BT 2004 [8] EXPERIMENT #2, LOW COMPLEXITY	VIDEO-TELECONFERENCE PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	ONE ORIGINAL 16 PVS 5-S CLIPS ORIGINALS WERE RATED	TELEVISION MONITOR (CRT) SPEAKERS	EMULATED BLOCKINESS VIDEO CONTENT: HEAD-AND-SHOULDERS	MNRU LEVEL THREE TO 21 AUDIO CONTENT: ■ SPEECH (MALE)	SINGLE STIMULUS (SSQS) FIVE-POINT SCALE
BT 2004 [8] EXPERIMENT #2, HIGH COMPLEXITY	VIDEO-TELECONFERENCE PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	ONE ORIGINAL 16 PVS 5-S CLIPS ORIGINALS WERE RATED	TELEVISION MONITOR (CRT) SPEAKERS	EMULATED BLOCKINESS VIDEO CONTENT: BICYCLE RACE	MNRU LEVEL THREE TO 21 AUDIO CONTENT: ■ SPEECH	SINGLE STIMULUS (SSQS) FIVE-POINT SCALE

NATIONAL UNIVERSITY OF SINGAPORE AND EPFL (2006) [9]	ENTERTAINMENT OVER 3GPP QCIF	PARTIAL MATRIX OF FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	SIX ORIGINALS 48 PVS ≈8-S ORIGINALS NOT RATED	PC MONITOR HEADPHONES	AVC FROM 24 TO 48 kb/S CODED AT 8 FPS ENTERTAINMENT CONTENT	MPEG-4 AAC-LC (LOW COMPLEXITY) FROM 8 TO 32 kb/S MONO AUDIO AUDIO CONTENTS: ■ ONE WITH SPEECH ■ ONE WITH MUSIC ■ FOUR MIXED	ACR 11-POINT SCALE
DEUTSCHE TELEKOM 2009 [10]	HDTV WITH PACKET LOSS IMPAIRMENT-FACTOR-BASED MODEL	PARTIAL MATRIX OF AUDIO-ONLY AND VIDEO-ONLY IMPAIRMENTS.	FIVE ORIGINALS 245 PVS 16-S CLIPS ORIGINALS WERE RATED.	PROFESSIONAL GRADE MONITOR (LCD) PROFESSIONAL GRADE SPEAKERS	AVC FROM TWO TO 16 Mb/S, WITH PACKET LOSS: FREEZING FROM 0% TO 0.25% AND SLICING FROM 0% TO 4%	MP2 FROM 48 TO 192 kb/S AND AAC AT 48 kb/S, WITH PACKET LOSS FROM 0% TO 8% AUDIO CONTENTS: ■ ONE WITH SPEECH ■ TWO WITH MUSIC ■ TWO MIXED	ACR 11-POINT SCALE, MAPPED TO 100-POINT SCALE
ITS (2009) [11]	ENTERTAINMENT TELEVISION CIF	TWO FULL MATRICES, EACH FOUR AUDIO ONLY BY FOUR VIDEO ONLY	TEN ORIGINALS 160 PVS 11–12-S CLIPS ORIGINALS WERE RATED	PC MONITOR PC SPEAKERS	H.263, VC-1, AVC, AND MPEG-2 FROM 75 TO 800 Kb/S ENTERTAINMENT CONTENT	M3, PCM, AND WMA FROM FOUR TO 32 kb/S MONO AUDIO AUDIO CONTENTS: ■ FIVE WITH SPEECH ■ THREE WITH MUSIC ■ TWO MIXED	ACR FIVE-POINT SCALE
ITS (2010)	ENTERTAINMENT TELEVISION HDTV	TWO FULL MATRICES, EACH FOUR AUDIO-ONLY BY FOUR VIDEO-ONLY SESSIONS HAD A RANDOM MIX OF AUDIO ONLY, VIDEO ONLY, AND AUDIO-VISUAL	TEN ORIGINALS 160 PVS 15-S EACH ORIGINALS WERE RATED	PROFESSIONAL GRADE TELEVISION MONITOR (LCD) PROFESSIONAL GRADE SPEAKERS	AVC FROM TWO TO SIX Mb/S AND MPEG-2 FROM SIX TO 12 Mb/S ENTERTAINMENT CONTENT	AAC FROM 16 TO 48 kb/S PLUS ORIGINAL STEREO AUDIO AUDIO CONTENTS: ■ TWO MUSIC ■ EIGHT SPEECH WITH MUSIC OR BACKGROUND NOISE	ACR FIVE-POINT SCALE

ln(***av***) rather than ***av***. This improved the Pearson correlation of their model number one from 0.956 to 0.980 [5].

META-ANALYSIS

Caution should be taken when vertically comparing the Pearson correlation (ρ) values in Table 2, because the denominator measures the range of quality within this particular experiment. Thus, while $\rho = 0.80$ would be poor for an experiment that contains a wide range of quality, $\rho = 0.80$ might be excellent for an experiment that spans a very limited range of quality (e.g., only high-quality video sequences suitable for broadcast television). Within each experiment (i.e., one horizontal row of the table), the correlation values and ranges can all be directly compared. (A purist might disagree. In some experiments, different viewer pools were used for audio only, video only, and audiovisual sessions.)

Likewise, when examining the range of MOS values spanned by each experiment, note the range of values available for that particular experiment (i.e., five-, nine-, 11-, or 100-point scale), the number of source sequences used (from one to ten), and the number of processed video sequences (PVSs) (from 16 to 245).

The oldest published studies were conducted by Bellcore from 1993 to 1995. These three separate experiments focused on standard definition television measured with a very small number of sequences (two) and artificial impairments. All three experiments reached the same conclusion: audio and video qualities are equally important in the overall audiovisual quality. Four later studies disagreed with Bellcore's conclusion: ITS 1998, KPN Research, Deutsche Telekom, and ITS 2009. These studies concluded that video quality was more influential than audio quality on the overall audiovisual quality. However, it can be observed that the range of quality spanned by audio is somewhat less than that spanned by video for these four experiments. When comparing max($\boldsymbol{a}$)-min($\boldsymbol{a}$) to max($\boldsymbol{v}$)- min($\boldsymbol{v}$), the audio range is 84%, 67%, 75%, and 47% of the video range, respectively. By contrast, the range of quality spanned by audio is similar to that spanned by video for the three Bellcore experiments (106%, 94%, and 97%, respectively). Thus, it is possible that this difference biased the experiments in favor of video. The three BT studies and the Singapore/EPFL study were inconclusive on this issue.

The ITS 2010 study described earlier was designed to test the hypothesis that, if the audio quality spanned nearly the same range as the video quality, then the audio and video quality would be equally important in the overall audiovisual quality. The range of quality spanned by audio was intended to be identical to that spanned by video, but ended up being slightly larger (113%). The ITS 2010 study agrees with Bellcore, concluding that audio and video have roughly the same influence on the overall audiovisual quality.

Bellcore also concluded that only the cross term ($\boldsymbol{a} \times \boldsymbol{v}$) is needed to predict the overall audiovisual quality. This conclusion is upheld by the generally stellar performance of this model in the other experiments conducted since then (see Model 1 in Table 2). Only two studies disagree with this

[TABLE 2] COMPARISON OF SUBJECTIVE AUDIOVISUAL MODELS FROM DIFFERENT LABORATORIES' EXPERIMENTS.

LABORATORY	MODEL	ρ	RANGE OF MOS	TYPE COMPARISON	DOMINANT FACTOR
BELLCORE 1993 [1]	1: $\hat{y} = 1.295 + 0.1077(\boldsymbol{a} \times \boldsymbol{v})$	0.99	$\boldsymbol{a}$ = [1.0 TO 8.2] $\boldsymbol{v}$ = [1.9 TO 8.7] $\boldsymbol{av}$ = [1.9 TO 8.3]	UNKNOWN	BOTH AUDIO AND VIDEO HAVE ROUGHLY THE SAME INFLUENCE
BELLCORE 1994 [2]	1: $\hat{y} = 1.07 + 0.1106(\boldsymbol{a} \times \boldsymbol{v})$	0.99	$\boldsymbol{a}$ = [1.4 TO 7.4] $\boldsymbol{v}$ = [1.5 TO 7.9] $\boldsymbol{av}$ = [1.8 TO 7.5]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.67 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.68 ρ	BOTH AUDIO AND VIDEO HAVE ROUGHLY THE SAME INFLUENCE
BELLCORE 1995 [3]	1: $\hat{y} = 1.912 + 0.114(\boldsymbol{a} \times \boldsymbol{v})$	0.99	$\boldsymbol{a}$ = [1.2 TO 7.3] $\boldsymbol{v}$ = [1.8 TO 8.1] $\boldsymbol{av}$ = [1.7 TO 7.3]	UNKNOWN	(MODEL CONSISTENT: ESSENTIALLY THE SAME AS [1] AND [2])
ITS (1998) [4]	1: $\hat{y} = 1.514 + 0.121(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -0.677 + 0.217a + 0.888\boldsymbol{v}$ 4: $\hat{y} = 0.517 - 0.0058\boldsymbol{a} + 0.654\boldsymbol{v} + 0.042(\boldsymbol{a} \times \boldsymbol{v})$	0.927 0.978 0.980	$\boldsymbol{a}$ = [1.5 TO 4.6] $\boldsymbol{v}$ = [1.0 TO 4.7] $\boldsymbol{av}$ = [1.1 TO 4.7]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.41 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.97 ρ $\boldsymbol{a}$ AND $\boldsymbol{v}$ = 0.29 ρ	VIDEO QUALITY
FRANCE TELECOM/CNET 1998 [5]	1: $\hat{y} = 1.76 + 0.10(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -0.13 + 0.35\boldsymbol{a} + 0.57\boldsymbol{v}$	0.960 0.956	$\boldsymbol{a}$ = [1.9 TO 4.5] $\boldsymbol{v}$ = [1.4 TO 4.8] $\boldsymbol{av}$ = [1.5 TO 4.9]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.42 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.86 ρ	COMPARED PASSIVE AND CONVERSATIONAL CONTEXT
KPN RESEARCH 1997 [6], [7]	1: $\hat{y} = 1.45 + 0.11(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = \alpha + \beta\,\boldsymbol{a} + \gamma\,\boldsymbol{v}$ 4: $\hat{y} = 1.12 + 0.007a + 0.24\boldsymbol{v} + 0.088(\boldsymbol{a} \times \boldsymbol{v})$	0.97 0.96 0.98	$\boldsymbol{a}$ = [3 TO 7] $\boldsymbol{v}$ = [2 TO 8] $\boldsymbol{av}$ = [2 TO 8]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.33 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.90 ρ	VIDEO QUALITY
BT 2004 [8] EXPERIMENT 1	1: $\hat{y} = \alpha + \mu(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = 4.26 + 0.59\boldsymbol{a} + 0.49\boldsymbol{v}$ 4: $\hat{y} = -3.34 + 0.85\boldsymbol{a} + 0.76\boldsymbol{v} + -0.01\,(\boldsymbol{a} \times \boldsymbol{v})$	0.72 0.97 0.99	$\boldsymbol{a}$ = [0 TO 63] $\boldsymbol{v}$ = [0 TO 71]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.74 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.62 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
BT [8] LOW COMPLEXITY	1: $\hat{y} = 1.15 + 0.17(\boldsymbol{a} \times \boldsymbol{v})$	0.85	$\boldsymbol{a}$ = [1.2 TO 4.8] $\boldsymbol{v}$ = [1.0 TO 4.6]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.61 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.55 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
BT [8] HIGH COMPLEXITY	1: $\hat{y} = \alpha + \mu\,(\boldsymbol{a} \times \boldsymbol{v})$ 3: $\hat{y} = 0.95 + 0.25\,\boldsymbol{v} + 0.15(\boldsymbol{a} \times \boldsymbol{v})$	0.79 0.85	$\boldsymbol{a}$ = [1.2 TO 3.8] $\boldsymbol{v}$ = [1.0 TO 4.3]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.44 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.68 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
NATIONAL UNIVERSITY OF SINGAPORE AND EPFL (2006) [9]	1: $\hat{y} = 1.98 + 0.103\,(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -1.51 + 0.456\boldsymbol{a} + 0.770\boldsymbol{v}$	0.94 0.94	$\boldsymbol{a}$ = [6 TO 9] $\boldsymbol{v}$ = [2 TO 8] $\boldsymbol{av}$ = [2 TO 8]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.55 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.67 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
DEUTSCHE TELEKOM 2009 [10] (SEE NOTE BELOW)	1: $\hat{y} = 30.917 + 0.007\,(\boldsymbol{a} \times \boldsymbol{v})$ 3: $\hat{y} = 27.805 + 0.129\boldsymbol{v} + 0.006\,(\boldsymbol{a} \times \boldsymbol{v})$	0.95 0.96	$\boldsymbol{a}$ = [30 TO 90] $\boldsymbol{v}$ = [20 TO 100] $\boldsymbol{av}$ = [30 TO 90]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.49 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.83 ρ	VIDEO QUALITY
ITS (2009) [11]	1: $\hat{y} = 1.1096 + 0.1959\,(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -0.5875 + 0.3599\boldsymbol{a} + 0.8037\boldsymbol{v}$ 4: $\hat{y} = 0.7500 - 0.0452\boldsymbol{a} + 0.3882\boldsymbol{v} + 0.1250\,(\boldsymbol{a} \times \boldsymbol{v})$	0.93 0.96 0.97	$\boldsymbol{a}$ = [2.3 TO 3.8] $\boldsymbol{v}$ = [1.3 TO 4.5] $\boldsymbol{av}$ = [1.0 TO 4.9]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.34 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.92 ρ	VIDEO QUALITY
ITS (2010)	1: $\hat{y} = 0.9616 + 0.1919\,(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -1.2757 + 0.6304\boldsymbol{a} + 0.6807\boldsymbol{v}$ 4: $\hat{y} = 0.9845 - 0.0525\boldsymbol{a} + 0.0274\boldsymbol{v} + 0.1969\,(\boldsymbol{a} \times \boldsymbol{v})$	0.96 0.94 0.96	$\boldsymbol{a}$ = [1.1 TO 4.6] $\boldsymbol{v}$ = [1.6 TO 4.7] $\boldsymbol{av}$ = [1.3 TO 4.8]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.68 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.66 ρ	BOTH AUDIO AND VIDEO HAVE ROUGHLY THE SAME INFLUENCE

NOTE: Information for Deutsche Telekom Model 1 was received in a private correspondence from Marie–Neige Garcia of Deutsche Telekom.

conclusion. The first is BT 2004, which shows a significant reduction in model accuracy when moving from the best model presented (additive) to the multiplicative model; and the second is BT High Complexity, which shows a moderate drop in correlation. It is possible that the very small number of video sequences used in these studies (two and one video sequences, respectively) resulted in measurement inaccuracies. Even so, BT concludes that people integrate audio and video errors together using a multiplicative rule, and that the true formula depends upon context and the test material under consideration [8]. Thus, in general, we see only a small drop in correlation when moving from the ideal model to the multiplicative model. The ITS 2010 study confirms this robust behavior of a model containing only the cross term.

While the other types of models (Models 2–4) have generally good performance, there is little agreement from one experiment to the next concerning the relative weight that should be assigned to β, γ, and μ. Some of these weights are very different indeed. For example, compare γ and μ for Model 2 as computed by ITS 1998 and BT Experiment 1. Likewise, there is no agreement as to which of these models is best (Models 2–4). This is problematic for someone who wishes to apply one of the other types of models, since it is unclear which set of weights should be chosen. The practical problem with the theory presented by BT (i.e., that different applications drive these differences) is that we do not have sufficient information available to say with any confidence the exact form of that model or the exact weights that are most appropriate for specific use case scenarios. Moreover, the accuracy gain for using one of the other models appears to be insignificant when the subjective experiment is designed with an approximately equal range of audio and video (compare Models 1–4 for ITS 2010).

CONCLUSIONS

There is no apparent pattern of relationship between the accuracy of the multiplicative model and the authors' conclusions as to the dominant factor (audio quality or video quality) on the overall audiovisual quality. This and the previous analyses (presented earlier) indicate that audio quality and video quality are equally important in the overall audiovisual quality. The application drives the range of audio quality and video quality examined and thus produces the appearance that one factor has greater influence than the other. The underlying perceptual model is invariant to application.

[AUDIO QUALITY AND VIDEO QUALITY ARE EQUALLY IMPORTANT IN THE OVERALL AUDIOVISUAL QUALITY.]

The most important overall conclusion is that only the cross term ($\boldsymbol{a} \times \boldsymbol{v}$) is needed to predict the overall audiovisual quality. It provides us with a simple and reasonably accurate model that has been tested in a wide variety of circumstances, from CIF to HDTV, from video conferencing to broadcast television, both coding only and with transmission errors, in a professional viewing/listening environment and on a PC. One missing factor is the impact of audiovisual synchronization errors (e.g., lip synchronization) on audiovisual quality. While many studies have been undertaken on audiovisual synchronization, further work is ongoing. A preliminary investigation on this topic undertaken by our lab is available in [16].

ACKNOWLEDGMENT

The ITS experiments presented herein were only possible due to contributions of video quality expertise from Stephen Wolf, audio quality expertise from Stephen Voran and Andrew Catellier, and computer programming from Scott Hanes.

Certain commercial equipment, materials, and/or programs are identified in this report to specify adequately the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration (NTIA), nor does it imply that the program or equipment identified is necessarily the best available for this application.

AUTHORS

Margaret H. Pinson (Margaret@its.bldrdoc.gov) received her B.S. and M.S. degrees in computer science from the University of Colorado at Boulder, in 1988 and 1990, respectively. Since 1988, she has been working as a computer engineer at ITS, an office of the NTIA in Boulder, Colorado. Her goal is to develop automated metrics for assessing the performance of video systems and actively transfer this technology to end users, standards bodies, and U.S. industry.

William Ingram (Bing@its.bldrdoc.gov) earned his B.S. and M.S. degrees in electrical engineering from Oklahoma State University in 1980 and 1981, respectively. Before joining NTIA/ITS in the late 1980s, he worked as an electrical engineer for the United States Department of Interior, Bureau of Reclamation. He has worked on a wide variety of projects at ITS, including the creation of automated compliance-testing systems for P25 radios, coauthored the Federal Standard "Glossary of Telecommunications Terms," and is currently working on several multimedia subjective testing projects.

Arthur Webster (Webster@its.bldrdoc.gov) received the B.A. degree in English and the M.S. degree in electrical engineering. He is the chair of ITU-T SG 9, "Television and Sound Transmission and Integrated Broadband Cable Networks." For many years, he has participated in technical standards groups, most of which are devoted to the standardization of video and multimedia quality assessment methods. He was a rapporteur (in SG9 or SG12) from 1995 to 2009 and has been the cochair of the Video Quality Experts Group since its founding in 1997. For the last 21 years, he has worked for the NTIA/ITS, where he manages two technical projects. He is a Member of the IEEE and ACM and holds two U.S. patents for innovations in objective assessment of video quality.

REFERENCES

[1] ANSI-Accredited Committee T1 Contribution, "Report on an experimental combined audio/video subjective test method," *Bellcore, T1A1.5/93-104*, Red Bank, New Jersey, July 22, 1993.

[2] ANSI-Accredited Committee T1 Contribution, "Report on extension of combined audio/video quality model," *Bellcore, T1A1.5/94-141*, Red Bank, New Jersey, July 22, 1993.

[3] ANSI-Accredited Committee T1 Contribution, "Combined A/V model with multiple audio and video impairments," *Bellcore, T1A1.5/94-124*, Red Bank, New Jersey, Apr. 10, 1995.

[4] C. Jones and D. Atkinson. (1998, May 18–20). *Development of opinion-based audiovisual quality models for desktop video-teleconferencing. Proc. Rec. 6th IEEE Int. Workshop Quality of Service, Napa, CA*. [Online]. Available: www.its.bldrdoc.gov/n3/video/documents.htm

[5] "Study of the influence of experimental context on the relationship between audio, video, and audiovisual subjective qualities," *ITU-T Contribution COM12-61-E*, France Telecom/CNET, France, Sept. 1998.

[6] "Relations between audio, video, and audiovisual quality," *KPN Res.,* The Netherlands, *ITU-T Contribution COM 12-19-E*, Feb. 1998.

[7] J. G. Beerends and F. E. de Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362, 1999.

[8] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, 2004.

[9] S.Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, 2006.

[10] M. N. Garcia and A. Raake, "Impairment-factor based audio-visual quality model for IPTV," in *Proc. Int. Workshop Quality of Multimedia Experience (QoMEx)*, 2009, pp. 1–6.

[11] M. McFarland, M. Pinson, C. Ford, A. Webster, W. Ingram, S. Hanes, and K. Anderson. (2009, Sept.). *Relating audio and video quality using CIF video. NTIA TM-10-472*. [Online]. Available: http://www.its.bldrdoc.gov/pub/ntia-rpt/10-472/

[12] Advanced video coding for generic audiovisual services. *ITU-T Recommendation H.264*, Geneva, Switzerland. [Online]. Available: http://www.itu.int/en/publications/Pages/default.aspx

[13] Advanced Video Coding, ISO/IEC 14496-10–MPEG-4 Part 10, Geneva, Switzerland. [Online]. Available: http://www.iso.org/iso/store.htm

[14] M. H. Pinson, S. Wolf, and G. Cermak. (2010, Mar.). *HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss. IEEE Trans. Broadcast.* [Online]. Available: www.its.bldrdoc.gov/n3/video/

[15] Modulated noise reference unit (MNRU). *ITU-T Recommendation P.810*, Geneva, Switzerland. [Online]. Available: http://www.itu.int/en/publications/Pages/default.aspx

[16] M. H. Pinson, A. Webster, and W. Ingram. (2011, Mar.). *Preliminary investigation into the impact of audiovisual synchronization of impaired audiovisual sequences. NTIA TM-11-474*. [Online]. Available: http://www.its.bldrdoc.gov/pub/ntia-rpt/11-474/

[17] S. Jumisko-Pyykkö and T. Vainio, "Framing the context of use for mobile HCI," *Int. J. Mobile Hum. Comput. Interact. (IJMHCI)*, vol. 2, no. 4, pp. 1–28, 2010.