“Summary of Subjective Audiovisual Quality Mapping”

Margaret H. Pinson
ITS Institute for Telecommunication Sciences

This VQEG contribution briefly summarizes 13 subjective experiments. Each of these experiments produced a model that mapped audio quality (***a***) and video quality (***v***) to the overall audiovisual quality (***av***). This information is presented to VQEG as an aid for discussions within the VQEG AVHD group.

Two tables below are taken from the following paper:

“Audiovisual Quality Components: An Analysis," *IEEE Signal Processing Magazine*, vol.28, no.6, pp.60-67, Nov. 2011, by Margaret H. Pinson, William Ingram, and Arthur Webster. Available at http://www.its.bldrdoc.gov/publications/2565.aspx

See that paper for the references to published information on each subjective experiment. 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: $\hat{y} = \alpha + \beta\ \boldsymbol{a} + \gamma\ \boldsymbol{v}$
3: $\hat{y} = \alpha + \gamma\ \boldsymbol{v} + \mu\ (\boldsymbol{a} \times \boldsymbol{v})$
4: $\hat{y} = \alpha + \beta\ \boldsymbol{a} + \gamma\ \boldsymbol{v} + \mu\ (\boldsymbol{a} \times \boldsymbol{v})$

A figure follows the tables.

TABLE 1

COMPARISON OF EXPERIMENT DESIGN FROM DIFFERENT LABORATORIES' INVESTIGATIONS INTO AUDIOVISUAL MODELS

Laboratory	Focus	Design	Size	Environment	Video	Audio	Scale
Bellcore 1993 **Error! Reference source not found.**	Entertainment television NTSC	Full matrix of 5 audio-only by 5 video-only. Originals not rated.	2 originals 50 PVS 18-sec 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)	9-point scale, (Excellent, good, fair, poor, unsatisfactory) similar to ACR
Bellcore 1994 **Error! Reference source not found.**	Entertainment television NTSC	Identical to Bellcore 1993 **Error! Reference source not found.**	Identical to Bellcore 1993 **Error! Reference source not found.**	Identical to Bellcore 1993 **Error! Reference source not found.**	Simulated blurring from a horizontal FIR low-pass filter Video content identical to Bellcore 1993 **Error! Reference source not found.**	Modulated Noise Reference Unit (MNRU) **Error! Reference source not found.** Audio content identical to Bellcore 1993 **Error! Reference source not found.**	Identical to Bellcore 1993 **Error! Reference source not found.**
Bellcore 1995 **Error! Reference source not found.**	Entertainment television NTSC	Three full matrices, each 3 audio-only by 3 video-only	2 originals 54 PVS 18-sec clips	Identical to Bellcore 1993 **Error! Reference source not found.**	Video impairments: • Same as **Error! Reference source not found.** • Same as **Error! Reference source not found.** • Simulated blockiness Video content identical to Bellcore 1993 **Error! Reference source not found.**	Audio impairments: • Same as **Error! Reference source not found.** • Same as **Error! Reference source not found.** • Random noise Audio content identical to Bellcore 1993 **Error! Reference source not found.**	Identical to Bellcore 1993 **Error! Reference source not found.**
ITS (1998) **Error! Reference source not found.**	Video-teleconference (VTC) NTSC	Audiovisual impairments created, and then split into audio-only and video-only Audiovisual bit rate 128 to 1536 kb/s.	6 originals 48 PVS 5 to 9-sec 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: • 6 with speech	ACR 5-point scale
France Telecom/CNET 1998 **Error! Reference source not found.**	Video-teleconference (VTC) PAL	Full matrix of 4 audio-only by 4 video-only.	2 originals 32 PVS 10-sec clips	Monitor and loudspeakers	Original, CIF and QCIF at 12 or 25fps from 172 to 456 kbps Video content: videoconference	Original, G.722,G.711, and G.728 from 16 to 56 kbps Audio content: • 2 with speech (1 male, 1 female)	ACR 5-point scale
KPN Research 1997 **Error! Reference source not found.**, **Error! Reference source not found.**	Broadcast television PAL	Full matrix 4 audio-only by 4 video-only.	2 originals 32 PVS 25-sec 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 9-point scale

Laboratory	Focus	Design	Size	Environment	Video	Audio	Scale
BT 2004 **Error! Reference source not found.** Experiment #1	Video-teleconference PAL	Full matrix 4 audio-only by 4 video-only.	2 originals 32 PVS 5-sec clips Originals were rated	Television monitor (CRT) Speakers	Emulated blockiness Video content: Low motion, low complexity	MNRU level 3 to 24 Audio content: • 2 with speech (1 male, 1 female)	DSCQS 100-point scale
BT 2004 **Error! Reference source not found.** Experiment #2, Low Complexity	Video-teleconference PAL	Full matrix 4 audio-only by 4 video-only.	1 original 16 PVS 5-sec clips Originals were rated	Television monitor (CRT) Speakers	Emulated blockiness Video content: Head-and-shoulders	MNRU level 3 to 21 Audio content: • speech (male)	Single stimulus (SSQS) 5-point scale
BT 2004 **Error! Reference source not found.** Experiment #2, High Complexity	Video-teleconference PAL	Full matrix 4 audio-only by 4 video-only.	1 originals 16 PVS 5-sec clips Originals were rated	Television monitor (CRT) Speakers	Emulated blockiness Video content: bicycle race	MNRU level 3 to 21 Audio content: • speech	Single stimulus (SSQS) 5-point scale
National University of Singapore & EPFL (2006) **Error! Reference source not found.**	Entertainment over 3GPP QCIF	Partial matrix of 4 audio-only by 4 video-only.	6 originals 48 PVS ≈8-sec Originals not rated	PC monitor Headphones	AVC from 24 to 48 kb/s coded at 8fps Entertainment content	MPEG-4 AAC-LC (low complexity) from 8 to 32 kb/s mono audio Audio contents: • 1 with speech • 1 with music • 4 mixed	ACR 11-point scale
Deutsche Telekom 2009 **Error! Reference source not found.**	HDTV with packet loss Impairment-factor-based model	Partial matrix of audio-only and video-only impairments.	5 originals 245 PVS 16-sec clips Originals were rated.	Professional grade monitor (LCD) Professional grade speakers	AVC from 2 to 16 Mbps, 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 kbps, with packet loss from 0% to 8% Audio contents: • 1 with speech • 2 with music • 2 mixed	ACR 11-point scale, mapped to 100-point scale
ITS (2009) **Error! Reference source not found.**	Entertainment television CIF	Two full matrices, each 4 audio-only by 4 video-only	10 originals 160 PVS 11 to 12-sec 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 & WMA from 4 to 32 kb/s mono audio Audio contents: • 5 with speech • 3 with music • 2 mixed	ACR 5-point scale
ITS (2010)	Entertainment television HDTV	Two full matrices, each 4 audio-only by 4 video-only Sessions had a random mix of audio-only, video-only, and audiovisual	10 originals 160 PVS 15-sec each Originals were rated	Professional grade television monitor (LCD) Professional grade speakers	AVC from 2 to 6 Mb/s and MPEG-2 from 6 to 12 Mb/s Entertainment content	AAC from 16 to 48 kb/s plus original stereo audio Audio contents: • 2 music • 8 speech with music or background noise	ACR 5-point scale

TABLE 2

COMPARISON OF SUBJECTIVE AUDIOVISUAL MODELS FROM DIFFERENT LABORATORIES' EXPERIMENTS

Laboratory	**Model**	**ρ**	**Range of MOS**	**Type Comparison**	**Dominant Factor**
Bellcore 1993 **Error! Reference source not found.**	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 **Error! Reference source not found.**	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 **Error! Reference source not found.**	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 **Error! Reference source not found.** and **Error! Reference source not found.**)
ITS (1998) **Error! Reference source not found.**	1: $\hat{y} = 1.514 + 0.121(\boldsymbol{a}\times\boldsymbol{v})$ 2: $\hat{y} = -0.677 + 0.217\boldsymbol{a} + 0.888\boldsymbol{v}$ 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 **Error! Reference source not found.**	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 **Error! Reference source not found.**, **Error! Reference source not found.**	1: $\hat{y} = 1.45 + 0.11(\boldsymbol{a}\times\boldsymbol{v})$ 2: $\hat{y} = \alpha + \beta\,\boldsymbol{a} + \gamma\,\boldsymbol{v}$ 4: $\hat{y} = 1.12 + 0.007\boldsymbol{a} + 0.24\boldsymbol{v} + 0.088(\boldsymbol{a}\times\boldsymbol{v})$	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 **Error! Reference source not found.** 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 **Error! Reference source not found.** 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 **Error! Reference source not found.**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 & EPFL (2006) **Error! Reference source not found.**	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[1] **Error! Reference source not found.**	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) **Error! Reference source not found.**	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

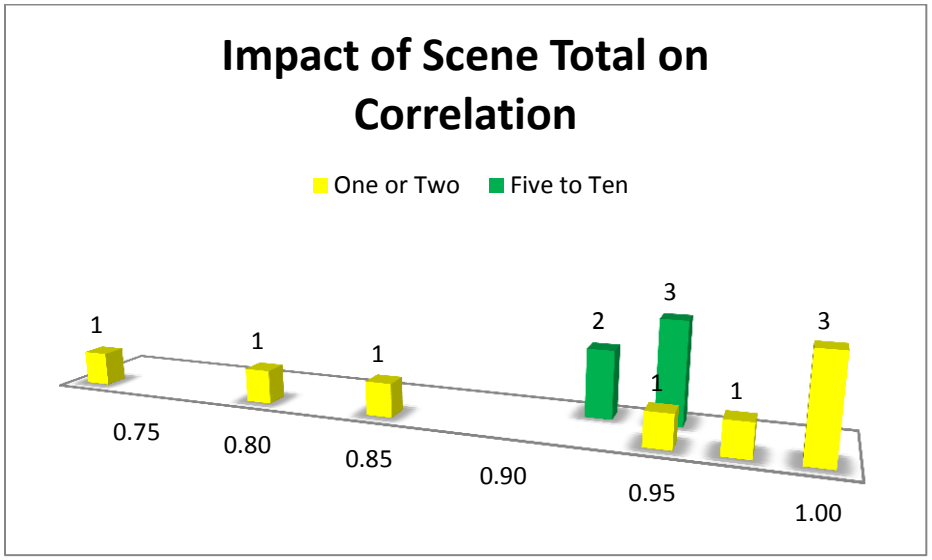

FIGURE 3.
HISTOGRAM OF THE PEARSON CORRELATION FROM 13 SUBJECTIVE EXPERIMENTS. EACH WAS DESIGNED TO ANSWER THE SAME QUESTION.

The above figure shows the accuracy of model type 1 (i.e., $\hat{y} = \alpha + \mu\ (\boldsymbol{a} \times \boldsymbol{v})$) for all 13 subjective experiments, based on the number of source sequences used. From this we can conclude that subjective tests with one or two source scenes were greatly influenced by chance. The subjective tests with five or more source sequences are tightly clustered, indicating a high degree of repeatability.

[1] Information for Deutsche Telekom model 1 was received in a private correspondence from Marie-Neige Garcia of Deutsche Telekom.

"Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP Journal on Image and Video Processing 2013*, 2013. Margaret H Pinson, Marcus Barkowsky, and Patrick Le Callet. Available at http://www.its.bldrdoc.gov/publications/2734.aspx