The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study — Source link

Margaret H. Pinson, Lucjan Janowski, Romuald Pepion, Quan Huynh-Thu ...+6 more authors

Institutions: AGH University of Science and Technology, Intel

Related papers:

- Study of Rating Scales for Subjective Quality Assessment of High-Definition Video

- A new standardized method for objectively measuring video quality

- Image quality assessment: from error visibility to structural similarity

- SOS: The MOS is not enough!

- Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies

The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study

Margaret Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Philip Corriveau, Audrey Younkin, Patrick Le Callet, William Ingram

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   1

The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study

Margaret H. Pinson[1], Lucjan Janowski[2], Romuald Pépion[3], Quan Huynh-Thu[4], Christian Schmidmer[5], Phillip Corriveau[6], Audrey Younkin[6], Patrick Le Callet[3], Marcus Barkowsky[3], William Ingram[1]

[1]National Telecommunications and Information Administration (NTIA), Boulder, CO 80305 USA

[2]Department of Telecommunication, AGH University of Science and Technology, Krakow, Poland

[3]LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597

(Institut de Recherche en Communications et Cybernétique de Nantes), Polytech Nantes, Nantes Cedex 3, France

[4]Technicolor Research & Innovation, Cesson-Sevigné, France

[5]OPTICOM, GmbH, Erlangen, Germany

[6]Intel Labs, Intel, Hillsboro, OR 97124, USA

Traditionally, audio quality and video quality are evaluated separately in subjective tests. Best practices within the quality assessment community were developed before many modern mobile audiovisual devices and services came into use, such as internet video, smart phones, tablets and connected televisions. These devices and services raise unique questions that require jointly evaluating both the audio and the video within a subjective test. However, audiovisual subjective testing is a relatively under-explored field. In this paper, we address the question of determining the most suitable way to conduct audiovisual subjective testing on a wide range of audiovisual quality.

Six laboratories from four countries conducted a systematic study of audiovisual subjective testing. The stimuli and scale were held constant across experiments and labs; only the environment of the subjective test was varied. Some subjective tests were conducted in controlled environments and some in public environments (a cafeteria, patio or hallway). The audiovisual stimuli spanned a wide range of quality.

Results show that these audiovisual subjective tests were highly repeatable from one laboratory and environment to the next. The number of subjects was the most important factor. Based on this experiment, 24 or more subjects are recommended for Absolute Category Rating (ACR) tests. In public environments, 35 subjects were required to obtain the same Student's t-test sensitivity. The second most important variable was individual differences between subjects. Other environmental factors had minimal impact, such as language, country, lighting, background noise, wall color, and monitor calibration. Analyses indicate that Mean Opinion Scores (MOS)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   2

are relative rather than absolute. Our analyses show that the results of experiments done in pristine, laboratory environments are highly representative of those devices in actual use, in a typical user environment.

Index Terms—Audiovisual Quality, Environment Effect, Language Effect, MOS, Subjective Testing.

I. INTRODUCTION

MODERN AUDIOVISUAL devices—tablets, laptops, smart phones, computers and connected TVs—raise new challenges for audio, video, and audiovisual subjective testing. How repeatable are audiovisual subjective tests from one laboratory to another? How critical are environmental constraints? For example, what if a quiet room is used instead of a sound isolation booth? How are audiovisual subjective scores impacted by conducting the experiment in a public location, such as a cafeteria?

There is an underlying philosophical question. International Telecommunication Union (ITU) Recommendations attempt to measure subjective quality in isolation. They assess subjective quality in an environment where the presentation device is considered to be transparent. This raises several questions. Do we want to measure quality perception in isolation? Do we want to measure quality perception in the environment where a device will be used? What is the impact of this choice?

This paper presents a set of audiovisual subjective experiments that were jointly conducted by six laboratories in four countries. The audiovisual test material and methodology were identical across all experiments. The test material included a wide range of audiovisual quality. However, the test environment was different among the experiments. Each laboratory conducted their experiments in either one or two environments of their choice. The primary goal was to study the impact of a laboratory environment versus a public environment on audiovisual quality subjective scores. The secondary goal was to identify non-controlled variables that need further study (e.g., lighting level, monitor size, objects on the wall, language, culture, vision).

Our analyses and conclusions focus on the impact of the number of subjects and test environment on subjective audiovisual quality ratings. Audiovisual subjective testing in public environments is desirable when answering complex questions, such as those posed in [1]-[2].

Other lines of research complement this effort, focusing on:

- Methodologies and rating scales [3]-[8]
- The number of response alternatives [9]
- Biases from the experiment design [10]
- The relative importance of audio quality and video quality in the perception of audiovisual quality [11]-[12]
- Quality of Experience (QoE) [13]
- Full length movies [2]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) < 3
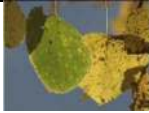
- Recency in instantaneous judgments [14]-[18]

- Forgiveness effect [19]-[20]

- Transmission errors [21]-[24]

- Visual attention [25]

II. EXPERIMENT DESIGN

This experiment used VGA resolution video (640x480) at 30fps. There were ten audiovisual source sequences for the test and one source sequence for the training session prior to the test. Each was 10 sec long (see Table 1). Five of the sequences contained speech or singing in English. Many of these sequences are available on the Consumer Digital Video Library (www.cdvl.org). The video was encoded in ITU-T H.264 [26], also known as Advanced Video Coding (AVC) [27]. Video coding bit-rates were 100, 192, 250, 448, 500, and 1000 kbps. The audio was encoded using Advanced Audio Coding (AAC) at 8, 32, and 64 kbps.

TABLE 1. SOURCE SEQUENCES

Sample Frame	Sequence Description
	Band "The Foot" music video, segment #2. Lyrics in English.
	Band "The Foot" music video, segment #8. Instruments only.
	Simulated news: commentator shows the Dushanbe Tea House. English speech plus background noise.
	Simulated news: reporter talking with cars driving in the background. English speech plus background noise. This content was used for the training session.
	"NTIA Halftime Music at Football Game" on www.cdvl.org Distant shot of a football half-time show. Instrumental music with crowd noise.
	"NTIA Speed Bag" on www.cdvl.org Boxer demonstrating punches on a speed bag. English speech plus punching.

Sample Frame	Sequence Description
	"NTIA Aspen Trees in Fall Color, Rapid Scene Cuts" on www.cdvl.org Instrumental music.
	"NTIA trio playing music, version 1, vga" on www.cdvl.org Non-professional musicians playing a lively piece of music. Used for training only.
	"Big Buck Bunny" from www.bigbuckbunny.org Animated sequence. Instrumental music with sound effects.
	Segment of "Elephant's Dream" from www.elephantsdream.org Animated sequence. Instrumental music with sound effects.
	"Big Green Rabbit (R) television revised open" from www.cdvl.org Animated sequence from a kids TV show. Lively music with singing in English.

The encoding levels were selected to avoid excessively unrealistic audio or visual impairments. However the audio bit-rates include levels that are lower than typically paired with the given video bit-rates. This was intentional, to ensure easily differentiated levels of audio quality and video quality that spanned similar ranges.

For each sequence, a high, medium, and low coding quality was selected for both audio and video. The experiment included the original source sequences plus five impaired versions of each source, for a total of 60 video clips. For each source, the five processed video sequences (PVSs) were chosen randomly from the nine possible combinations. So, for some PVSs the audio quality was higher than the video quality; and for some PVSs the video quality was higher than the audio quality. This yielded a wide range of audiovisual quality, but prohibits some types of statistical analysis (e.g., analysis of variance (ANOVA) to determine the relative importance of audio quality versus video quality).

The training session used four versions of another video clip: the original plus three impaired versions. Subjects were provided with written instructions and heard detailed instruction during training that described the task to be performed. Subjects were asked to watch/listen to a series of audio-video sequences, and rate the overall (combined audio-video) quality of each sequence on the absolute category rating (ACR) scale [42], [43]. ACR was implemented as a five-point discrete quality rating scale (ratings: Excellent, Good, Fair, Poor and Bad). With this approach, each PVS is presented one at a time and rated independently.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   5

The order of presentation of the PVSs was randomized between participants. Rating was not time-limited. Subjects were asked to judge the audio-video reproduction quality and not the quality of the program content. The laboratories differed slightly in training technique (e.g., wording of instructions, answers to subject questions).

All experiments were conducted on a computer (desktop or laptop) and used the same software, which was a Java interface that played the video and saved the subjective ratings. All videos were played at their native resolution and at the native resolution of the display (no up-sampling to full-screen). To ensure correct playback on a variety of different computers, the 60 PVSs were very lightly compressed into Windows Media Video format (WMV). A variable rate coder was used, resulting in compression bit-rates from 5 Mb/s to 15 Mb/s, depending upon the compression difficulty. All subjects saw the same compressed WMV files.

Each experiment was conducted in one of the following environments:

- A controlled environment that meets the spirit of a pristine environment from ITU-R BT.500 [29], ITU-T P.910 [42] and/or ITU-R P.800 [43]. These controlled environments did not necessarily meet the ITU specifications.
- A public location, such as a cafeteria, patio or hallway. The public environments had other people talking and going about their own business.

The experiment was conducted in ten different environments, to produce ten different datasets. These are summarized in Table 2 to Table 5. The viewing distance is expressed in both picture heights and angular degree (i.e., the approximate angle that VGA picture spanned at the subject's eye). Fig. 1 illustrates two of the test environments.

TABLE 2. DATASET DESCRIPTION: LABORATORY, LANGUAGE AND DEVICES

Dataset #	Lab	Native Language	English Proficiency	Device
1	NTIA	English	Native	Broadcast quality monitor (LCD)
2	NTIA	English	Native	Laptop
3	Intel	English	Native	Professional LCD
4	IRCCyN	French	Beginner to advanced	Professional LCD
5	IRCCyN	French	Beginner to advanced	Tablet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   6

6	Technicolor	French	Intermediate to advanced	Good monitor
7	Technicolor	French	Intermediate to advanced	Laptop
8	AGH	Polish	None to advanced	Laptop
9	AGH	Polish	None to advanced	Laptop
10	OPTICOM	German	Beginner or intermediate; three advanced	Good monitor

TABLE 3. DATASET DESCRIPTION: SCREEN TECHNICAL DETAILS AND VIEWING DISTANCE

#	Size	Screen Resolution	View Distance	Screen Brightness	Screen Color
1	24"	1920x1080	≈6H, 8º	Default	Calibrated
2	17"	1920x1200	≈4H, 11º	Default	Default
3	42"	1920x1080	≈3H, 20º	Calibrated	Calibrated
4	40"	1920x1080	≈6H, 8º	Calibrated	Calibrated
5	7"	1024x600	≈4H, 13º	Maximum	Default
6	19"	1280x1024	≈6H, 10º	50%	Default
7	15"	1920x1200	≈6H, 10º	Maximum	Default
8	15"	1400x1050	≈4H, 13º	Maximum	Default
9	15"	1400x1050	≈4H, 13º	Maximum	Default
10	24"	1920x1200	≈4H, 13º	Calibrated (100 cd/m^2)	Calibrated

The cafeteria used for dataset 2 closed unexpectedly, so only nine subjects are available. Datasets 4 and 5 used the same subjects. Half of these subjects rated dataset 4 and then dataset 5, while the other half rated dataset 5 and then dataset 4. Datasets 6 and 7 used 18 of the same subjects. Dataset 6 was conducted one week before dataset 7.

Ethical approval for experimentation on human subjects was obtained. Parental permission was attained for all minors. Subjects were not required to know English. Subjects for NTIA and AGH were paid, temporary workers. Subjects for Opticom were uncompensated volunteers. Subjects for Intel and IRCCyN were associated with these large organizations and thanked for their participation with movie coupons. Subjects for Technicolor were unpaid volunteers who were interested in participating.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   7

TABLE 4. DATASET DESCRIPTION: LIGHTING LEVEL AND ENVIRONMENT

Dataset Number	Audio Playback	Lux	Environment Notes
1	Speakers	25	Laboratory: sound isolation chamber, grey walls, adjustable lighting.
2	Earbuds	150+	Public cafeteria with varying numbers of people talking. Indirect sunlight and fluorescent lighting.
3	Speakers	20	Laboratory: semi-anechoic chamber, grey walls
4	Head-phones	200	Laboratory: grey furniture and walls, background (of the display) illumination.
5	Head-phones	150+	Public cafeteria with varying numbers of people talking. Indirect sunlight.
6	Head-phones	20	Laboratory: black walls, 20 lux background illumination, very quiet
7	Head-phones	150+	Patio with varying numbers of people talking. Bright sunlight.
8	Head-phones	200	Laboratory
9	Head-phones	150+	Hallway with a few people walking past. Indirect sunlight.
10	Head-phones	120	Home office with proper lighting. Quiet environment

TABLE 5. DATASET DESCRIPTION: INFORMATION ON SUBJECTS

Dataset Number	Total Subjects	Vision not 20/20	Color Blind	Expert Viewers	Subject Ages
1	28	0	2	0	18-65
2	9	0	0	0	18-65
3	34	2	1	0	21-56
4	25	0	0	0	18-46
5	25	0	0	0	18-46
6	24	0	0	0	25-57

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   8

7	24	0	0	0	25-57
8	14	1	0	0	18-56
9	15	0	0	0	18-56
10	15	0	0	2	<14 or >40

Fig. 1. Two sample environments. Top shows dataset 1; bottom shows dataset 2.

Subjects from datasets 8 and 9 saw the English ACR scale on the monitor (excellent, good, fair, poor, bad) but were given a sheet of paper with the Polish translation (bardzo dobry, dobry, średni, słaby, zły) and asked to use that mapping instead of relying upon their English skill. Similarly subjects from datasets 6 and 7 were provided with a written French translation they could refer to (Excellent, Bien, Satisfaisant, Médiocre, Mauvais) although they used the English labels to provide their ratings during the test. Subjects from dataset 10 were verbally instructed to ensure they understood the German meanings of the English ACR scale. The training instructions included some small differences (e.g., datasets 6 and 7 mentioned that the video would not be full-screen and that this was intentional; while datasets 8 and 9 did not).

III. DATA ANALYSIS

A. Subject Correlations and Eliminating Invalid Subjects

Fig. 2 shows the correlation between each subject's ratings and the Mean Opinion Score (MOS) of that dataset. Dataset 3 shows the highest correlations. This indicates that all subjects made their judgments using the same criteria. Wider distributions could indicate a bimodal distribution; that is subjects may have made judgments using different criteria. This is the reason to use a panel of subjects and take the mean. Fig. 2 uses red squares to highlight datasets performed in public environments.
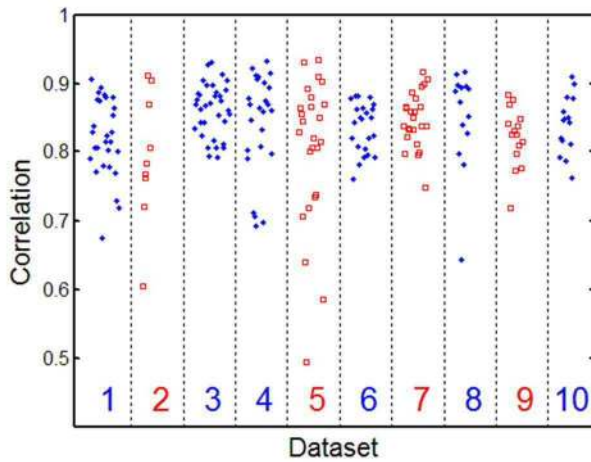
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   9

Fig. 2. Correlation between each subject's votes and dataset MOS. Public environments are red squares (2, 5, 7 & 9); controlled environments are blue dots (1, 3, 4, 6, 8, & 10).

The observed differences may be due to outlier subjects. Post-hoc screening of subjective results is usually conducted to detect and eliminate subject outliers. The problem is to determine an appropriate method to eliminate subjects who are inattentive or confused without also eliminating subjects with genuine differences of opinion.

The standard algorithms eliminate outliers based on score distributions (e.g., kurtosis, correlation). The problem is that we are not sure why people are responding differently. For example, the differences might result from English audio when a subject does not speak any English. Thus, we took a conservative approach by considering everyone's judgments as valid. We decided to eliminate only subjects who did not understand the task. This differs from standard practices in subjective testing [28]-[30].

One subject was eliminated from dataset 9 due to extremely low correlation (0.09). This subject was interviewed after scoring. The notes confirmed that this subject did not understand the task: he was rating the production quality (opinion of the program) instead of the video quality (reproduction). This subject is in addition to the 15 subjects mentioned in Table 5.

Subjects who failed the color vision test or the 20/20 vision test were retained (see Table 5). In dataset 10, two expert viewers were mixed with naïve subjects. Hearing was not tested. None of these subjects were outliers. For example, the subject-to-dataset correlations for the two colorblind subjects in dataset 1 were 0.77 and 0.83 (see Fig. 2).

B. What Matters Most: The Number of Subjects

This section examines how Pearson correlation is impacted by the number of subjects in a dataset. Fig. 3 shows a histogram of the global Mean Opinion Scores (MOS_G). For these purposes, MOS_G was computed as an average of all subjects' ratings combined across all datasets. The MOS_G ranges from 1.05 to 4.58. This figure shows that the design goal was met: the stimuli span a wide range of audiovisual quality.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   10
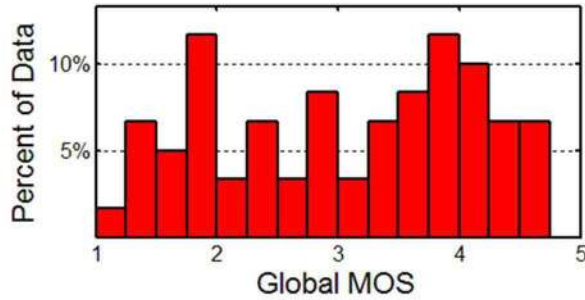
Fig. 3. Histogram shows the distribution of MOS_G. Bins are 0.25 wide.

Table 6 shows the Pearson correlation between each pair of datasets. The number of subjects in each dataset is listed in the bottom row of the table (labeled "#"). Fig. 4 presents these same correlations in a histogram. The correlations range from 0.93 to 0.99. Pearson correlation is calculated using the MOS of all 60 video clips, across a pair of datasets. These high correlations indicate that language, lighting, background noise, monitor calibration, and other environmental factors had minimal impact on the quality difference between the stimuli. The impact of personal opinion, language, and environment will be further discussed in sections III.K, **Error! Reference source not found.**, and III.M

The other important pattern we see is that the number of subjects is critical. Dataset 2 has only nine subjects and its correlation with other datasets ranges from 0.93 to 0.96. Datasets 1 and 3–7 have 24 to 33 subjects. Dataset-to-dataset correlations within this group range from 0.96 to 0.99. "Number of subjects" was the most important control variable for a repeatable subjective experiment.

TABLE 6. DATASET-TO-DATASET CORRELATIONS, WITH NUMBER OF SUBJECTS

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.95	0.98	0.97	0.97	0.97	0.98	0.97	0.96	0.95
2	0.95	1.00	0.95	0.94	0.94	0.93	0.96	0.94	0.93	0.93
3	0.98	0.95	1.00	0.98	0.98	0.98	0.99	0.98	0.97	0.97
4	0.97	0.94	0.98	1.00	0.98	0.96	0.97	0.97	0.96	0.96
5	0.97	0.94	0.98	0.98	1.00	0.96	0.97	0.96	0.97	0.96
6	0.97	0.93	0.98	0.96	0.96	1.00	0.99	0.97	0.97	0.95
7	0.98	0.96	0.99	0.97	0.97	0.99	1.00	0.97	0.97	0.96
8	0.97	0.94	0.98	0.97	0.96	0.97	0.97	1.00	0.96	0.96
9	0.96	0.93	0.97	0.96	0.97	0.97	0.97	0.96	1.00	0.97
10	0.95	0.93	0.97	0.96	0.96	0.95	0.96	0.96	0.97	1.00

#	28	9	34	25	25	24	24	14	15	15

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) < 11
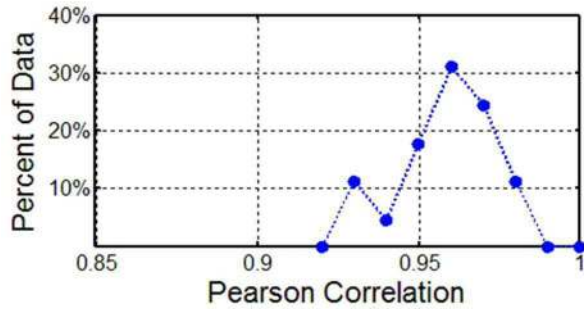
Fig. 4. Dataset-to-dataset correlations, plotted as a histogram. This figure was computed using all subjects and datasets.

C. How Correlation Drops with Fewer Subjects

The problem with Fig. 4 is that the different datasets contain different numbers of subjects. The next two figures use subsets of the available data to indicate trends.

Fig. 5 shows the impact of the number of subjects in an experiment on dataset-to-dataset correlation. This figure uses only the six datasets that have 24 or more subjects: datasets 1 and 3–7. Five subjects were chosen at random from each of the six datasets, and the dataset-to-dataset correlations computed. This process was repeated for nine, 15, and 24 subjects. The subsets of subjects were chosen randomly, and the results averaged over 500 runs to ensure stability. The plotted lines show dataset-to-dataset correlations for 24, 15, 9, and 5 subjects.

The dataset-to-dataset correlation curves improve as more subjects are used. From what the data shows, smaller sets of subjects can be used for pilot data, since it will predict trending. The smaller datasets can be followed up by the more statistically reliable testing with 24 or more subjects.

The choice to use 15 subjects is of particular interest as this number is recommended by international recommendations such as [29]. The 15 subject recommendation is supported by [31], which presents analysis using a combination of simulated data and five unrelated subjective datasets. In Fig. 5, the 24 subject curve clearly indicates a much more stable and repeatable experiment result across different test labs. This averaged dataset-to-dataset correlation is always 0.96 or greater, which indicates a well conducted experiment.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   12
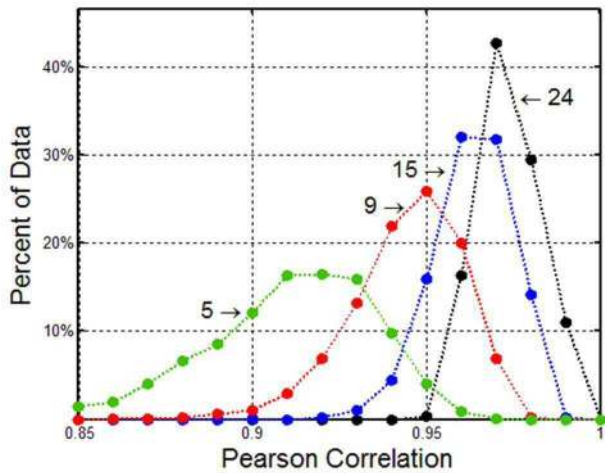
Fig. 5. Number of subjects in a dataset and the impact on dataset-to-dataset correlations. Plotted for 24, 15, 9 and 5 subjects (see labels on the plot).

D. How Correlation Drops with a Narrow Range of Quality

Let us now take an opposing approach, and consider the impact of the range quality on dataset-to-dataset correlations. This range results from the experiment design. Where this experiment contains a wide range of quality, other experiments might be designed to span a narrower range of quality. To understand this relationship, a deeper understanding of Pearson correlation is required.

Pearson correlation takes the form:

$$\rho^2 = 1 - \frac{\alpha}{\beta} \qquad\qquad (1)$$

where ρ is correlation [44]. Roughly stated, **β** measures the overall spread of MOS in datasets *A* and *B*; and **α** measures, for each individual PVS, the distance between the MOS of dataset *A* and dataset *B*. The variable **α** captures the differences of opinions that occur when two different pools of subjects rate the same sequence. These differences of opinion do not go away when subjects are shown a narrow range of video quality.

Fig. 6 shows the relationship between the range quality and the dataset-to-dataset correlations. Fig. 6 uses the eight datasets with 15 or more subjects. For each curve, 15 subjects were randomly chosen and results were averaged over 500 runs. The curve on the right (black circles) shows the histogram of dataset-to-dataset correlations using all clips. These clips span a range of quality that is 3.58 MOS units wide (from 1.05 to 4.58). The middle curves (solid lines) use clips that span a range of quality that is 3.0 MOS units wide. The left curves (dashed lines) use clips that span a range of quality that is 2.5 MOS units wide. The curves with red up-pointing triangles, green squares and blue down-pointing triangles retain clips toward the top, middle and bottom of the quality scale, respectively. Notice that the curves in Fig. 6 are not impacted by whether we chose good quality clips, poor quality clips, or supposedly "difficult to score" clips (i.e., avoiding the ends of the scale). Correlation decreases as the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   13

range of quality decreases.

Why does this occur? Refer back to Equation (1). When a subjective experiment spans a small range of video quality, τ shrinks, ε stays about the same, and Pearson correlation drops. Statistical analyses can be misleading, as Huff demonstrates in *How to Lie with Statistics* [32].

The goal of this section is not to discourage the use of Pearson correlation; rather, it is to put the high correlations seen in Fig. 2 and Table 6 into perspective. Those high correlations are possible because this was a well-designed experiment that contained a wide range of audiovisual quality and several large groups of subjects.
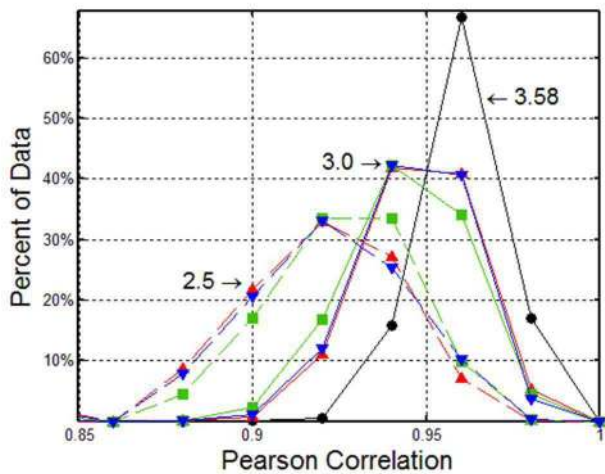
Fig. 6. The range of MOS in a dataset and its impact on dataset-to-dataset correlations.

E. *Subset Analysis: Splitting a Dataset into Two Pieces*

The question then arises, how can we tell whether or not our correlations are good? To that end, we will split each dataset into two randomly chosen subsets and correlate them. This allows us to compare correlations within a dataset to correlations between datasets.

Given dataset A of size $Length_A$, and dataset B of size $Length_B$, calculate N:

$$N = \left\lfloor \frac{\min(\quad_A, \quad_B)}{2} \right\rfloor \qquad (2)$$

Let us draw from A two disjoint sets A_1 and A_2, each containing N clips. Let us draw from B two disjoint sets B_1 and B_2, each containing N clips. We will then compute:

$$\rho_A = \qquad\quad_1, A_2)$$
$$\rho_B = \qquad\quad_1, B_2) \qquad (3)$$

Let us then draw set A_3 from A and set B_3 from B, each containing N clips. Sets A_3 and B_3 are drawn independently from A_1, A_2, B_1 and B_2. Compute:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   14

$$\rho_{AB} = \qquad\qquad 3, B_3) \qquad\qquad (4)$$

This procedure was repeated 5000 times, and plotted as histograms. Fig. 7 shows six of the 45 plots. Since N is the same for all three curves in each plot, sample size does not influence the conclusions.
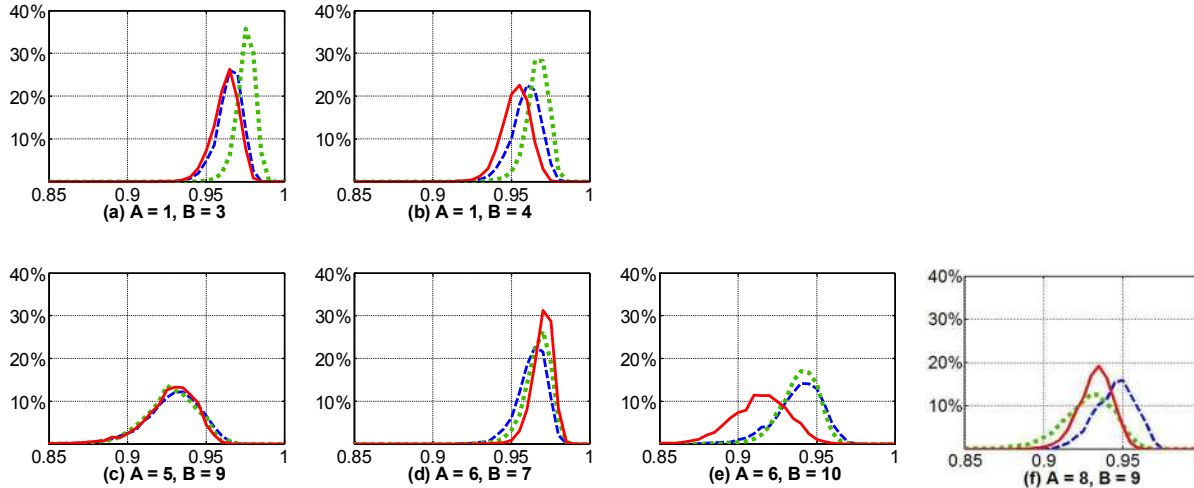

Fig. 7. . Histograms show the range of ρ_A, ρ_B, and ρ_{AB}, for example dataset pairs. The blue dashed line is ρ_A, the green dotted line is ρ_B, and the red solid line is ρ_{AB}. The x-axis is correlation, and the y-axis is frequency. Sub-plots show dataset pairs: (a) A=1 & B=3, (b) A=1 & B=4, (c) A=5 & B=9, (d) A=6 & B=7, (e) A=6 & B=10, (f) A=8 & B=9.

For example, Fig. 7 (a) compares dataset 3 with dataset 1. ρ_1 (the blue, dashed line) shows 14 subjects from dataset 1 correlated with another 14 subjects from dataset 1. ρ_3 (the green, dotted line) shows 14 subjects from dataset 3 correlated with a different 14 subjects from dataset 3. ρ_{13} (the solid, red line) shows 14 subjects from dataset 1 correlated with 14 subjects from dataset 3. ρ_3 is ≈ 0.97, and ρ_1 is ≈ 0.95. Thus, we can see that dataset 3 subjects agree with each other more than dataset 1 subjects.

Similarly, we can conclude that subjects' ratings in datasets 5 and 9 are equally spread. The most common behavior is that ρ_{AB} is a bit worse than ρ_A or ρ_B (see Fig. 7 (a), 1 vs. 3, and Fig. 7 (b), 1 vs. 4). Sometimes ρ_{AB}, ρ_A and ρ_B are all very similar (see Fig. 7 (c), 5 vs. 9).

Sometimes ρ_A and ρ_B are very similar yet ρ_{AB} is a much worse (see Fig. 7 (e), 6 vs. 10). Results indicate some differences between the datasets. However, the exact reasons cannot be identified at this point. The data collected in Table 2 to Table 5 do not explain these differences. We will see this again in later analyses.

One possible explanation is simply that different subjects were chosen. Nevertheless, it means that almost all subjects from one group are different from subjects in the other group.

In rare cases ρ_{AB} is a bit better than ρ_B (see Fig. 7 (f), 8 vs. 9) or better than both ρ_A and ρ_B (see Fig. 7 (d), 6 vs. 7). This can be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   15

explained by outliers (see section III.A: subjects with differing opinions were retained in all ten datasets), as these impact the computation of correlations in Equations (3) and (4). Outliers are chosen for subsets A_1 or A_2 with a high probability, because A_1 and A_2 when combined typically contain most of the subjects in dataset A. The probability that the outlier is part of A_3 is considerably less.

F. Kruskal-Wallis Test: Comparing Absolute MOS

Our analysis now touches upon a philosophical question. Are MOS absolute or relative? When the results of two subjective experiments are statistically different, the reason can be difficult to explain. We accept that the results of these experiments are different. Are we willing to accept these differences as a consequence of different subject behavior, even if all other factors are fixed? Can all dataset-to-dataset differences could be explained if we gather sufficient data? Or is there a flaw in our choice of statistic?

The ratings in a subjective experiment are ordinal. That is, we know the order but we do not know the distance between ratings. For simplicity, most analyses assume equal distances between two consecutive scores on the rating scale. This assumption has been questioned. [33] and [34] recommend the use of different statistics and ordinal values.

The Kruskal-Wallis test does not make any assumption about the distance between rating values. Kruskal-Wallis is a non-parametric test of the hypothesis that all groups have the same median. If it is not true, Kruskal-Wallis pairwise comparisons can be run. If each category on our scale has an absolute meaning, then the Kruskal-Wallis test can be used to identify statistically different datasets. Note that the Kruskal-Wallis test does not assume that subjects cannot behave differently. It compares the differences within each group with the differences among groups. If the differences among groups are larger than those within group, it clearly shows that differences among groups come from factors other than differences observed among subjects for each group.

Fig. 8 shows Kruskal-Wallis pairwise comparisons. These are based only on the mean rank and the number of answers for each dataset. Therefore, datasets with fewer subjects have much wider confidence intervals. Notice that:

- All datasets do not have the same median.
- Datasets 1, 3, 4, 8 and 9 have the same median. These datasets were conducted in Boulder, Colorado (USA); Portland, Oregon (USA); and Krakow (Poland).
- Datasets 5 and 6 have the most dissimilar medians. These datasets were conducted in Nantes (France) and Rennes (France). The difference could be explained by the fact that only dataset 5 used a tablet (see Table 2).

Therefore, the Kruskal-Wallis test indicates that language and country did not have a statistically significant impact on these datasets, for the languages used in these experiments. This can be seen in Fig. 8 by considering the issues mentioned above. No

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   16

tonal language was considered.

When a dataset has fewer than 24 subjects, statistically significant differences are difficult to find. For example, datasets 4 and 5 are statistically different, and datasets 8 and 9 are statistically indistinguishable. The distances between pairs of medians is approximately the same. If datasets 4 and 5 had 15 subjects (like datasets 8 and 9), we would probably not be able to make such conclusions. (Note that datasets 4 and 5 used the same subjects.)
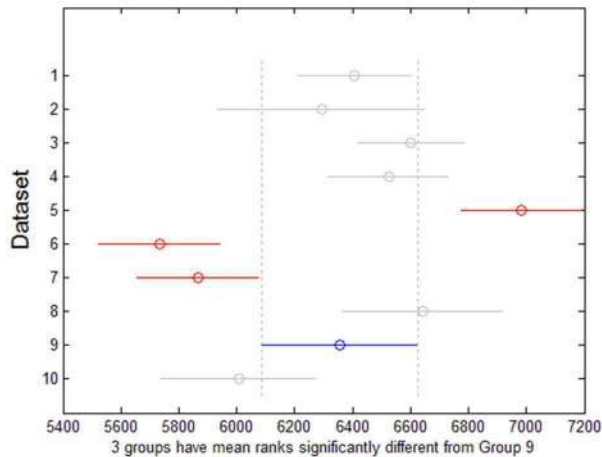

Fig. 8. Kruskal-Wallis confidence intervals for all datasets. Dataset 9 selected for pairwise comparison. Dataset 9 is closest to the global mean.

The Kruskal-Wallis test can be used to look for statistically significant differences between dataset responses to an individual source sequence or PVS. A separation of audio and video impairments is difficult, as this experiment did not contain a full matrix of those conditions.

Both source and PVS analyses show that the differences observed at the data set level are not caused by a single sequence or distortion. Rather, these analyses show fewer significant differences within datasets than between datasets. This means that differences between datasets come from accumulating small differences that are seen but are not statistically significant at the source or PVS level.

G.   Is MOS Relative?

Another school of thought on subjective data is that MOS are always relative. In other words, the ordering of impairments is consistent, but absolute quality ratings are not in general repeatable. This claim is made in [35] using subjective data from the VQEG Full-Reference Television Test Phase I.

To explore this idea, let us scale each dataset's MOS. All subjects' scores from all datasets were averaged into a global Mean Opinion Score (MOS_G), as shown on Fig. 3. The linear fit between each dataset and this global MOS_G is shown in Table 7.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   17

TABLE 7. LINEAR MODEL FIT BETWEEN EACH DATASET AND GLOBAL MOS

Dataset	Linear Fit to MOS_G	Correlation
1	$\hat{y} = 0.96\ x_1 +\ \ 0.11$	0.99
2	$\hat{y} = 1.00\ x_2 +\ \ 0.02$	0.96
3	$\hat{y} = 0.97\ x_3 +\ \ 0.01$	1.00
4	$\hat{y} = 1.00\ x_4 +\ \ -0.05$	0.99
5	$\hat{y} = 1.05\ x_5 +\ \ -0.39$	0.99
6	$\hat{y} = 0.94\ x_6 +\ \ 0.40$	0.99
7	$\hat{y} = 0.93\ x_7 +\ \ 0.39$	0.99
8	$\hat{y} = 0.95\ x_8 +\ \ 0.07$	0.98
9	$\hat{y} = 1.03\ x_9 +\ \ -0.08$	0.98
10	$\hat{y} = 0.93\ x_{10} +\ \ 0.34$	0.98

Dataset 5 has, on average, MOS scores 0.47 higher than dataset 6—yet correlation is minimally impacted (see Table 6). This supports the theory that MOS are relative.

What are the consequences, if MOS are relative? A fit is needed before two datasets can be compared. A linear fit seems sufficient for this experiment. Another consequence is that the same subjects should be used when comparing different factors (e.g., two screens, two devices).

Also, some statistical tests may be inappropriate. Pearson correlation is not impacted by a linear scaling between datasets. However, the Kruskal-Wallis Test may be unsuitable, because a shift in the median is unimportant.

H.   Confidence Interval (CI) Analysis: Controlled Environments have Slightly Better CI

Fig. 9 shows the 95% confidence interval (CI) for each dataset. This gives us some insights into the differences between controlled environments and public environments. CI drops as the number of subjects increases, as a function of 1/sqrt(N), so the x-axis is plotted on a logarithmic scale. These CI were calculated using random subsets of subjects, run repeatedly and then averaged.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   18
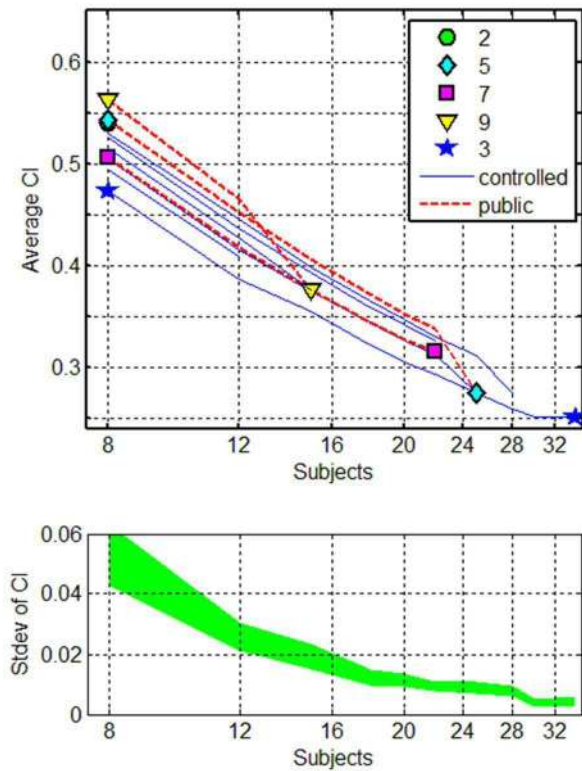
Fig. 9. Relationship between the 95% confidence interval (CI) of each dataset and the number of subjects. The x-axis plots the number of subjects on a logarithmic scale (base 10). The top plots the CI for each dataset on the y-axis. The bottom shows the standard deviations for these CI averages.

Fig. 9 indicates a modest increase in CI when moving from a controlled environment to a public environment. We can compensate for a public environment by using more subjects in the test. However, as MOS seems to be relative (see section III.G), identical CI do not guarantee that an identical fraction of PVS pairs can be differentiated.

I. Paired Comparisons: Compensate for a Public Environment with Extra Subjects

Usually, subjective experiments are designed to compare different impairments. For example, we may wish to compare the performance of two transmission error concealment techniques (such as [36]) or investigate the impact of coding bit-rate on 3DTV quality (such as [37]).

Often, the Student's t-Test is applied, to evaluate whether or not the quality differences are statistically significant. Does a controlled environment allow us to distinguish between more conditions than a public environment? Certainly, the public environments have the potential for more noise and distractions. Do the controlled environments that more closely follow existing ITU Recommendations allow the best discrimination?

Fig. 10 presents the results of the Student's t-Test conducted for each dataset. Note that the "number of subjects" was switched from the x-axis to the y-axis, to better display the data. Each unique pair of PVSs was compared, and the percent of pairs that could be distinguished computed. Random subsets were chosen for different numbers of subjects (8, 12, 15, 18, 20, 22, 25, 28,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   19

30 and 34). The plot shows the average result from repeated trials.
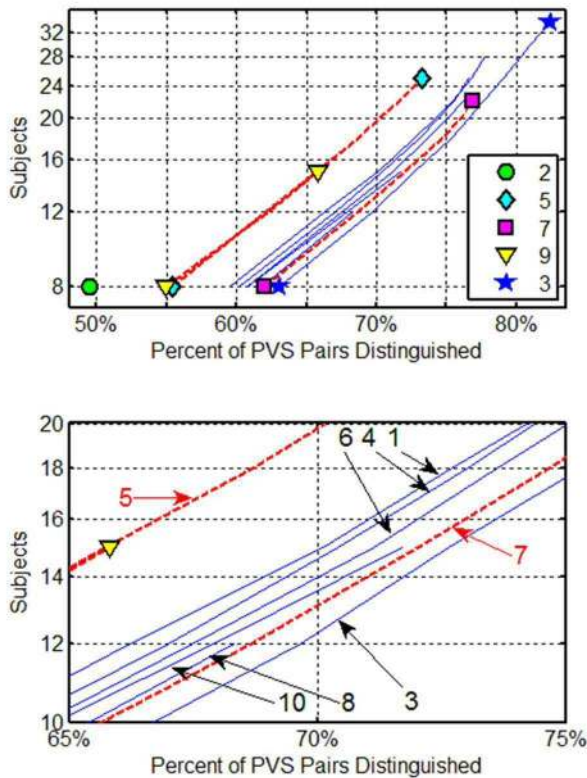

Fig. 10. Relationship between the discrimination power of each dataset and the number of subjects. The y-axis plots the number of subjects on a logarithmic scale (base 10). The top graph shows relative distribution of controlled environments (thin blue) and public environments (thick dashed red). The bottom graph zooms in on 70% with 14 subjects.

As the number of subjects increases, the percent of PVSs that can be distinguished increases. However, we see diminishing returns as more subjects are added (notice that the y-axis is plotted on a logarithmic scale). We do not see a pattern among controlled environment datasets; there is no evidence that a dataset's performance improves with closer adherence to existing ITU Recommendations. Rather, all controlled environments have very similar performance (see the blue lines in Fig. 10).

Comparing public and controlled environments, we see mixed results. Datasets 2, 5 and 9 (public environments) show worse behavior than datasets 1, 3, 4, 8 and 10 (controlled environments). Dataset 7 (public environment) had a discrimination power similar to or better than every experiment except dataset 3. This is a similar pattern to that seen in Fig. 5.

This phenomenon can be counteracted by using a larger set of subjects. Fifteen subjects in a controlled environment seem to perform similarly to 22 subjects in a public environment. By extrapolation, we would expect 24 subjects in a controlled environment to perform similarly to ≈35 subjects in a public environment. This is a larger increase in subjects than we saw with CI. Datasets 5 and 9 were used for these calculations, as they provide the median response.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   20

J. Repeatability: How Often do Datasets Reach Different Conclusions?

Are the same pairs of PVS responsible for these statistically significant differences we see in Fig. 10? The problem in answering that question is this: if we choose one value for the "percent of PVS pairs distinguished," we get a different number of subjects for each dataset. Thus, we cannot compute a unique set of PVSs that were responsible for the differences between dataset pairs—the answer depends upon the subset of subjects selected.

We can instead count the number of agreements between different subjective experiments. Let us fix the fraction of PVS pairs distinguished at 73%. There are different numbers of subjects for each dataset as shown in Table 8. We will not consider dataset 5 in this analysis, because 25 subjects would be required. This is all of the subjects available in dataset 5, so multiple random subsets cannot be chosen.

TABLE 8. NUMBER OF SUBJECTS NEEDED TO DISTINGUISH 73% OF PVS PAIRS

Dataset	1	3	4	6	7
Subjects	19	15	19	18	17

For each test, the above number of subjects is used to compute the Student's t-test between all possible PVS pairs. Using random subsets of subjects, the Student's t-tests were repeated 100 times. This yields a percent chance that a particular PVS pair was distinguished, for each dataset. We then take the absolute difference between the percentages for a pair of datasets on a PVS pair-by-pair basis, and calculate average over all PVS pairs. In other words, we calculate the chance of the following occurrence: one dataset was able to distinguish between PVS pairs and the other dataset was not (see Table 9). This ignores the chance that the two datasets were both able to distinguish between the PVS pairs but reached opposite conclusions.

All values are less than 16%, so the results may be considered consistent across the datasets in general. Still, this 16% may concern exactly the sequences that are of interest.

TABLE 9. PERCENT OF TRIALS WHERE ONE DATASET DISTINGUISHED BETWEEN PVS PAIRS, AND THE OTHER DATASET DID NOT

	1	3	4	6	7
1		14%	15%	16%	13%
3	14%		12%	11%	10%
4	15%	12%		16%	13%
6	16%	11%	16%		10%
7	13%	10%	13%	10%	

Also of interest is the chance that the two datasets were both able to distinguish between the PVS pairs but reached opposite

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

\> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   21

conclusions. For fewer than 18 observers, this will occasionally occur. However, even with 8 observers, the probability was less than 0.03%.

K. The Impact of Personal Opinion on Experiment Design

The number of subjects was the most important variable that impacted dataset-to-dataset correlations. The second most important variable was differences of opinion from one person to another. It is possible that demographic information could explain some of the differences between subjects that we saw in Section III.E.

Experiment design should reflect the importance of subject-to-subject differences. The analyses presented above indicate a 24 subject minimum in a controlled environment, and 35 in a public environment. Smaller sets of subjects can be used for pilot tests, to find trends. The smaller experiments can be followed up by the more statistically reliable testing with 24 or more subjects.

Twenty-four or more subjects were insufficient to guarantee identical PVS rankings from one dataset to another, when statistical significance was computed.

L. The Impact of Language on the Quality Scale

Several speech quality experiments have performed similar comparisons using the ACR five-point scale and a set of impairments. Cai [38] compared Chinese, Japanese and English and found lab-to-lab correlations ranging from 0.903 to 0.95. Goodman [39] compared Britain, Canada, France, Italy, Norway and USA and found lab-to-lab correlations ranging from 0.919 to 0.985. As is standard procedure in speech quality, different speech samples were used by each laboratory (i.e., speech samples from several native speakers of that language).

By contrast, our ten datasets used identical audio samples. Our datasets were collected in different countries with different native languages. Some of the stimuli were in English. Nonetheless, the results were similar and the test repeatable. Language and country did not have a statistically significant impact on these datasets (see Section III.F).

This was our most interesting and surprising result.

Part of the explanation lies in research presented in [40]. Words used to label subjective scales have different meanings in different languages. On the surface, it seems that the ACR scale should thus have nonlinearities. In [40] three audio tests performed on the same stimuli are described: one with ACR, one with MUSHRA [45] and one with an unlabeled scale. The dataset-to-dataset correlations were all 0.99. "One possible explanation is that the listeners ignored the meaning of any labels and used the graphic scales without reference to the labels, or perhaps only taking the end point labels into account." [40]

M. Pristine Environments

Let us consider the limitations of the experiment described in this paper. The stimuli spanned a wide range of quality—both

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   22

seen and heard. By contrast, ITU-R BS.1116 [41] is intended for high quality audio. Likewise, ITU-R BT.500 is targeted at analyzing differences between high quality video sequences. The pristine environments specified in these Recommendations allow subjects to perceive very small differences in quality. The experiment described in this paper is more similar to the ITU-T P.910 scenario [42].

ITU-R BT.500 and ITU-R BS.1116 use very tightly constrained and controlled environments. Our analyses show that experiments done in pristine environments are highly representative of those devices in actual use, in a typical user environment.

IV. CONCLUSION

Do we want to measure quality perception in isolation? Do we want to measure quality perception in the environment where a device will be used? The impact of this choice appears to be minimal, when testing a wide range of audiovisual quality. Quality perception measured in isolation accurately predicts quality perception in the environment where a device will be used.

The total number of subjects appears to be the most important control variable. Our study indicates that 24 or more subjects should be used when in a controlled environment. It is recommended to increase to 35 subjects when using a public environment or a narrow range of audiovisual quality. Smaller numbers of subjects are suitable for pilot studies, to find trending.

The second variable is people—how opinions differ among subjects. Subjects drawn from any one source cannot fully replicate the behavior of "all people". Because diverse opinions are so important, improved methods are needed for eliminating non-performing subjects. Current methods assume opinions are homogenous (e.g., Appendix V of [28]).

MOS appears to be relative and not absolute. In other words, we expect the ordering of impairments and relative distances to be replicable, so statistics such as Pearson correlation are appropriate. Other statistical tests may be inappropriate (e.g., those that depend upon the mean or median being correct, or comparison to a constant MOS threshold). If different factors are to be compared, the same subjects should be used and the factors (e.g., two screens, two devices) should be presented in random order.

Device / display may have had a relative impact. The use of a tablet in dataset 5 might be responsible for the differences seen in Fig. 5 and section III.H. The device, monitor and viewing distance may have more of an impact on scores if the experiment is explicitly designed to test this variable.

The following factors did not seem to matter—or at least mattered so little that the difference was obscured by human factors:

- Native language / speech comprehension

- Culture / country of origin

- Lighting

- Background noise

- Wall color

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   23

- Objects on the wall

- Viewing distance

- Monitor calibration

- Color blindness

- Vision good but not 20/20

- Translation of ACR scale labels

These constraints had significantly less impact on the subjective test results than differences between subjects, yet are more likely to be controlled and reported today.

The statistical power of our study is limited by the large number of non-controlled variables. Ideally, smaller yet more controlled follow-on tests should be conducted by a variety of researchers. For a small added cost, the same pool of subjects can be used to score an experiment in two different environments, as was done by Catellier in [46].

Environmental factors should not always be ignored. Some environmental factors may have a strong influence, depending upon the experiment design and purpose. For example, eliminating background noise is critical when subjects are expected to detect subtle sound differences. Monitor calibration is important when evaluating studio quality monitors. Language is important for a comprehension test or an audiovisual synchronization test. The importance of background noise may increase if subjects used speakers instead of headphones or earbuds.

Subject screening is an issue that deserves further consideration. This experiment did not justify discarding subjects for slightly imperfect hearing, slightly imperfect vision, color blindness, or being an outlier in opinion score distribution. However, some screening is clearly needed to eliminate people who are inattentive or do not understand the task. The influence of training and instructions is likewise an area for further research.

The impact of language and culture on subjective scores would be an interesting topic for further investigation. In the video quality community many researchers claim that "cultural and language differences result in statistically significant differences obtained for the same experiment run in different countries." While these ten experiments appeared not to be influenced by language or culture, we do not have enough data to fully explain or support this conclusion for the general case. Ideally, a precise experiment would focus on this issue directly (e.g., identical sequences rated by two labs per country, using nearly identical environments and equipment).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   24

thank Thomas François for his assistance in data collection.

REFERENCES

[1] F. N. Rahayu *et al.*, "Subjective video quality assessment in the presence of audio for digital cinema," *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, Sep. 2011.

[2] N. Staelens, *et al.*, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *IEEE Transactions on Broadcasting*, vol. 56, no. 4, Dec. 2010.

[3] Q. Huynh-Thu *et al.*, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Transactions on Broadcasting*, vol. 57. no. 1, p. 1-14, Mar. 2011.

[4] S. Péchard *et al.*, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm." *Proc. Int. Workshop on Image Quality and its Application, Kyoto, Sep.* 2008.

[5] M. Brotherton *et al.*, "Subjective multimedia quality assessment*," IEICE Trans. Fundamentals, Electron. Commun. Comput. Sci.*, vol. E89-A, no. 11, pp. 2920–2932, 2006.

[6] T. Tominaga *et al.*, "Performance comparisons of subjective quality assessment methods for mobile video," *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, Jun. 2010.

[7] P. Corriveau *et al.*, "All subjective scales are not created equal: the effects of context on different scales," *Signal Processing*, vol. 77, no. 1, pp1-9, 1999.

[8] Q. Huynh-Thu and M. Ghanbari, "A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video," *Proc. of Signal and Image Processing*, Honolulu, Hawaii, USA, vol. 479, 2005.

[9] E. P. Cox, "The optimal number of response alternatives for a scale: a review," *Journal of Marketing Research*, vol. XVII, Nov., 1980, pp. 407-422.

[10] S. Zielinski *et al.*, "On some biases encountered in modern audio quality listening tests—a review," *Journal of Audio Engineering Societ*y, vol. 56, no 6, Jun. 2008.

[11] M. Pinson *et al.*, "Audiovisual quality components," *IEEE Signal Processing Magazine*, Nov. 2011.

[12] J. G. Beerends *et al.*, "The influence of video quality on perceived audio quality and vice versa," *Journal of Audio Engineering Society*, vol. 47, no. 5, May 1999.

[13] G. Cermak, "Four suggestions for research on multimedia QoE using subjective evaluations," *IEEE Communications Society Multimedia Communications Technical Committee (COMSOC MMTC) E-Letter*, vol. 4 no. 10, Nov. 2009. Available online: http://committees.comsoc.org/mmc/eletters.asp

[14] D. E. Pearson, "Viewer response to time-varying video quality", *Proc. SPIE Human Vision and Electronic Imaging III*, vol. 3299, 26-29 Jan., 1998.

[15] R. Aldridge R *et al,*. "Measurement of scene-dependent quality variations in digitally coded television pictures", *IEE Proceedings: Vision, Image and Signal Processing*, vol.142, no.3, Jun. 1995.

[16] R. Hamburg and H. De Ridder, "Time-varying image quality: modeling the relation between instantaneous and overall quality," *SMPTE Journal,* vol. 108, no. 11, Nov. 1999.

[17] D. S. Hands and S. E. Avons, "Recency and duration neglect in subjective assessment of television picture quality." *Applied Cognitive Psychology*, vol. 15, no 6, Nov. 2001.

[18] L. Gros and N. Chateau, "Instantaneous and overall judgements for time-varying speech quality: assessments and relationships," *Acta Acustica united with Acustica*, vol. 87, no. 3, May-June 2001.

[19] V. Seferidis *et al.*, "Forgiveness effect in subjective assessment of packet video", *Electronics Letters*, vol. 28, no. 21, Oct. 1992.

[20] D. S. Hands, "Temporal characterisation of forgiveness effect", *Electronics Letters*, vol. 37, no. 12, Jun. 2001.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   25

[21] Y. Liang *et al.*, "Analysis of packet loss for compressed video: does burst-length matter?" *Proc. of IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2003.

[22] R. R. Pastrana-Vidal *et al.*, "Sporadic frame dropping impact on quality perception," *Proc. SPIE Human Vision and Electronics Imaging IX*, vol. 5292, Jan. 2004, pp. 182-193.

[23] R. R. Pastrana *et al.*, Temporal masking effect on dropped frames at video scene cuts," *Proc. SPIE Human Vision and Electronic Imaging IX*, vol. 5292, Jan. 2004, pp. 194-201.

[24] U. Reiter *et al.,* "Comparing apples and oranges: assessment of the relative video quality in the presence of different types of distortions," *EURASIP Journal on Image and Video Processing*, 2011.

[25] O. Le Meur and P. Le Callet, "What we see most is likely to be what matters: visual attention and applications," *Proc. IEEE International Conference on Image Processing*, Cairo, Egypt, 2009.

[26] Recommendation ITU-T H.264, (2011). *Advanced video coding for generic audiovisual services*, Geneva, 2011. Available on-line: http://www.itu.int

[27] ISO/IEC 14496-10:2010*, Information technology -- Coding of audio-visual objects – Part 10: Advanced Video Coding*.

[28] M. Pinson *et al.*, "Report on the validation of video quality models for high definition video content," VQEG, 2010. Available: http://www.vqeg.org.

[29] Recommendation ITU-R BT.500-12, *Methodology for the subjective assessment of the quality of television pictures*, Geneva, 2009. Available on-line: http://www.itu.int

[30] Recommendation ITU-T P.78 (1996), *Subjective testing method for determination of loudness ratings in accordance with Recommendation P.76*, Geneva, 1996. Available on-line: http://www.itu.int

[31] S. Winkler, "On the properties of subjective ratings in video quality experiments," *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX) 2009*, 29-31 Jul. 2009.

[32] D. Huff and I. Geis, *How to lie with statistics*, W. W. Norton & Company, New York, N.Y., 1993.

[33] A. Głowacz *et al.*, "Automated qualitative assessment of multi-modal distortions in digital images based on GLZ," *Special Issue of Annals of Telecommunications on Quality of Experience and Socio-Economic Issues of Network-Based Services*, Springer, vol. 65, no. 1-2, pp. 3-17, Feb. 2010.

[34] L. Janowski, and Z. Papir, "Modeling subjective tests of quality of experience with a generalized linear model", *Quality of Multimedia Experience (QoMEX)*, Jul. 2009.

[35] M. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," *Proc. SPIE Video Communications and Image Processing Conference*, Lugano, Jul. 2003.

[36] Y. Pitrey *et al.*, "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," *SPIE Optical Engineering*, San Diego, USA, Aug. 2010.

[37] K. Wang *et al.*, "Subjective evaluation of HDTV stereoscopic videos in IPTV scenarios using absolute category rating," *Proc. SPIE Stereoscopic Displays and Applications XXII*, San Francisco, Jan. 2011.

[38] Z. Cai *et al.*, "Comparison of MOS evaluation characteristics for Chinese, Japanese, and English in IP Telephony," *4th International Universal Communication Symposium (IUCS)*, Oct. 2010.

[39] D. J. Goodman and R. D. Nash, "Subjective quality of the same speech transmission conditions in seven different countries," *IEEE Transactions on Communications*, vol. 30, no. 4, 1982.

[40] S. Zieliński *et al.*, "On the Use of Graphic Scales in Modern Listening Tests," *Proc. 123rd Convention of the Audio Engineering Society*, Oct. 2007.

[41] Recommendation ITU-R BS.1116-1, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, Geneva, 1997. Available on-line: http://www.itu.int

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

\> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   26

[42] Recommendation   ITU-T P.910, *Subjective video quality assessment methods for multimedia applications*, Geneva, 2008. Available on-line:
http://www.itu.int

[43] Recommendation ITU-T P.800, *Methods for subjective determination of transmission quality,* Geneva, 1996. Available on-line: http://www.itu.int

[44] J. Neter *et al.*. "Applied linear statistical models," McGraw-Hill Companies, Inc., 1974, p81.

[45] Recommendation ITU-R BS.1534-1, *Method for the subjective assessment of intermediate quality levels of coding systems,* Geneva 2003. Available on-line: http://www.itu.int

[46] A. Catellier *et. al.*, "Impact of mobile devices and usage location on perceived multimedia quality," *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, Jul. 2012.

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, USA, in 1988 and 1990, respectively.

Since 1988, she has been working as a researcher at the Institute for Telecommunication Sciences (ITS) in Boulder, Colorado. Margaret H. Pinson joined the Video Quality Program in 1989. She is project leader of the Video Quality Research Program. 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. She is a Co-Chair of the Independent Lab Group (ILG) in the Video Quality Experts Group (VQEG) and an Associate Rapporteur of Question 12 in ITU-T Study Group 9. She has served as technical editor for VQEG Final Reports and ITU-T Recommendations. She administers the Consumer Digital Video Library.

Lucjan Janowski received his M.Sc. degree in Telecommunications in 2002 and Ph.D. degree in Telecommunications in 2006 both from the AGH University of Science and Technology.

Since 2006 he has worked at AGH University of Science and Technology, where he is currently an assistant professor in the Department of Telecommunications. During 2007 he held a post-doc position in CNRS-LAAS (Laboratory for Analysis and Architecture of Systems of CNRS) in France, where he prepared both malicious traffic analysis and anomaly detection algorithms. In 2010/2011 he spent half a year on a post-doc position in University of Geneva working on QoE for health applications. He is QoE team member at AGH. Lucjan Janowski is also co-chair of the JEG-Hybrid group within the Video Quality Experts Group (www.vqeg.org).

Quan   Huynh-Thu received the Dipl.-Ing. degree in electrical engineering from the University of Liège (Belgium), the M.Eng. degree in electronics engineering from the University of Electro-Communications (Japan) and the Ph.D. degree in electronic systems engineering from the University of Essex (UK).

He is currently Senior Scientist at Technicolor. His main research interests for the past 12 years have been related to image processing and human visual perception, with a focus on perceptual visual quality, visual attention, human factors and user experience. He was Research Scientist in the Image and Signal Processing Lab at the Belgian Forensic Institute from 1997 to 2000. Being awarded a fellowship from the Japanese Ministry of Education, he joined the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <　　27

University of Electro-Communications, Japan, as a Researcher from 2000 to 2003. He was Senior Research Engineer with Psytechnics Ltd, UK, from 2003 to 2010, where he co-developed a perceptual video quality metric included in the ITU-T Recommendation J.247 for the objective measurement of video quality.

Since 2004, he has been actively contributing to the work of standards-related bodies addressing video quality, including the International Telecommunication Union (ITU) and the Video Quality Experts Group (VQEG). He is currently Rapporteur for Question 2 in ITU-T Study Group 9, and co-chair of both the VQEG 3DTV and Multimedia groups.

Christian Schmidmer studied electronic engineering at the University of Erlangen. After achieving his M.S. degree (Diplom) he spent five years as a scientist at the audio department of the famous Fraunhofer Institute for Integrated Circuits in Erlangen (the home of mp3), mostly dedicated to the research of psychoacoustics and the development of perceptual measurement tools as well as audio codecs, contributing to the development of mp3.

In 1997 he joined OPTICOM as CTO and co-owner. OPTICOM's core business is the development and IPR management for voice, audio and video quality measurement algorithms. Christian Schmidmer is active in standardization bodies like ITU, VQEG and ETSI. He is the author of many scientific publications and frequently presenting papers at conferences and workshops. He is one of the main developers behind the recommendations ITU-R BS.1387 / PEAQ (Perceptual Evaluation of Audio Quality), ITU-T P.563 / 3SQM (no-reference voice quality assessment) and ITU-T P.863 / POLQA (full reference voice quality assessment).

Philip Corriveau is a Principal Engineer in the technology arm of the Interaction & Experience Research group in Intel Labs. Philip received his Bachelors of Science Honors at Carleton University, Ottawa Canada in 1990. He immediately started his career at the Canadian Government Communications Research Center performing end-user subjective testing in support of the ATSC HD standard for North America. In January 2009 he was awarded a National Academy of Television Arts & Science, Technology & Engineering Emmy® Award for User Experience Research for the Standardization of the ATSC Digital System.

Philip moved to Intel in 2001 to seed a research capability called the Media and Acoustics Perception Laboratory designed to address fundamental perceptual aspects of platform and product design. He now manages a team of human factors engineers in the Experience Metrics & Quality Evaluation group conducting user experience research across Intel technologies, platforms and product lines.

Philip is currently the Vice Chair for 3D@Home and International 3D Society driving the research addressing Human Factors issues surrounding the development of 3D technologies for end-users. He was a founding member and still participates in the Video Quality Experts Group, aimed a developing, testing and recommending for standardization objective video quality metrics.

Audrey Younkin is a perceptually focused Senior Human Factors Engineer with Intel's Labs-Interaction and Experience Research. She received a BS in Biology from the University of Portland, Portland, OR USA, in 2003.

She has worked in the audio-visual industry for the last seven years. Her current research incorporates human perception of video and audio in order to understand thresholds across different platforms and devices. Her e-mail is audrey.c.younkin@ intel.com.

Patrick Le Callet received M.Sc. degree PhD degree in image processing from Ecole polytechnique de l'université de Nantes. He was also student at the Ecole Normale Superieure de Cachan where he got the "Aggrégation" (credentialing exam) in electronics of the French National Education. He has working as an Assistant professor from 1997 to 1999 and as a full time lecturer from 1999 to 2003 at the department of Electrical engineering of Technical Institute of University of Nantes (IUT). Since 2003 is teaching at Ecole polytechnique de l'université de Nantes (Engineer School) in the Electrical Engineering and the Computer Science department where is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

> FOR CONFERENCE-RELATED PAPERS, REPLACE THIS LINE WITH YOUR SESSION NUMBER, *E.G.*, AB-02 (DOUBLE-CLICK HERE) <   28

now Full Professor. Since 2006, he is the head of the Image and VideoCommunication lab at CNRS IRCCyN, an academic research group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. His current centers of interest are 3D image and video quality assessment, watermarking techniques and visual attention modeling and applications. He is co-author of more than 190 publications and communications and co-inventor of 13 international patents on these topics. He has coordinated and is currently managing for IRCCyN several National or European collaborative research programs representing grants of more than 2,5 million euros. He is serving in VQEG (Video Quality Expert Group) where he is co-chairing the "HDR Group" and "3DTV" activities. He is currently serving as associate editor for IEEE transactions on Circuit System and Video Technology, SPIE Journal of Electronic Imaging and SPRINGER EURASIP Journal on Image and Video Processing.

Marcus Barkowsky received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. Starting from a deep knowledge of video coding algorithms his Ph.D. thesis focused on a reliable video quality measure for low bitrate scenarios. Special emphasis on mobile transmission led to the introduction of a visual quality measurement framework for combined spatio-temporal processing with special emphasis on the influence of transmission errors. He joined the Image and Video Communications Group IRCCyN/IVC at the University of Nantes in 2008, and was promoted to associate professor in 2010. His activities range from modeling effects of the human visual system, in particular the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays using psychometric and medical measurements.

William Ingram graduated from Oklahoma State University in 1981 with a Bachelor's degree in Electrical Engineering and then earned his Master's degree in Electrical Engineering 1982.

William was hired as an Electronics Engineer with the U.S. Department of Interior in 1983. He then transferred to the U.S. Department of Commerce two years later and has worked at the Institute for Telecommunication Sciences since then.