SUBJECTIVE AND OBJECTIVE EVALUATION OF AN AUDIOVISUAL SUBJECTIVE DATASET FOR RESEARCH AND DEVELOPMENT

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

[1] National Telecommunications and Information Administration (NTIA), [2] Intel Labs, Intel
[3] OPTICOM, [4] Department of Telecommunication, AGH University of Science and Technology,
[5] LUNAM Université, Université de Nantes, IRCCyN, [6] Technicolor R&D France

ABSTRACT

In 2011, the Video Quality Experts Group (VQEG) ran subjects through the same audiovisual subjective test at six different international laboratories. That small dataset is now publically available for research and development purposes.

Index Terms**—** audiovisual, subjective testing

1. INTRODUCTION

In 2011, the Video Quality Experts Group (VQEG) performed a subjective audiovisual quality test [1]. The test dataset can be found on the Consumer Digital Video Library (CDVL, www.cdvl.org) by searching for the keyword "vqegMM2". In this test, 189 people rated simple distortions that were distributed over a wide range of perceptual quality. This dataset is not critical enough to differentiate between reasonably accurate objective models, but vqegMM2 can be used to develop subject screening algorithms, design a larger audiovisual test, pretest an audiovisual metric, or compare the same subjective test run in different labs, countries, and conditions. If the ratings from all labs are combined, then the influence of number of ratings on mean opinion score (MOS) can be analyzed.

2. OVERVIEW OF THE VQEGMM2 DATASET

In 2011, six laboratories from four countries conducted a systematic study of audiovisual subjective testing. The goal was to explore the impact of environment and laboratory on audiovisual subjective scores. The stimuli and scale were held constant across experiments and labs; only the environment of the subjective test was varied. Analyses show that these audiovisual MOSs were highly repeatable from one lab and environment to the next. See [1] for a complete description of this dataset's design and an analysis of the subjective scores.

This dataset is nicknamed "vqegMM2", since it was conducted under VQEG's Multimedia Phase II (MM2) project. That effort has since been moved into VQEG's new audiovisual HD project (AVHD). VqegMM2 contains ten source sequences at VGA resolution video (640 × 480), 30 fps, and 10 seconds long. Eight of the sequences contained music or singing in English, and two contain speech with background noise. Table I characterizes each source (SRC) using spatial information (SI) and temporal information (TI) from ITU-T P.910. SI and TI measure the amount of fine detail and motion in the SRC, respectively. Source 8 is missing from Table I, because it was used for training the subjects while they were introduced to their task.

Table I. SI and TI from ITU-T P.910

SRC	1	2	3	4	5	6	7	9	10	11
SI	**152**	**98**	**68**	**54**	**89**	**77**	**89**	**76**	**87**	**89**
TI	**69**	**43**	**27**	**41**	**78**	**21**	**48**	**27**	**51**	**9**

VqegMM2 does not use a full matrix experiment design. The intent was an even distribution of quality, not an analysis of bitrate. Audio and video impairments were separately chosen to span similar ranges of quality. The encoding used Advanced Audio Coding (AAC) and ITU-T H.264, respectively. For each source sequence, a high, medium, and low video coding quality level was manually selected, and five processed video sequences (PVS) were chosen randomly from the nine possible combinations (i.e., three audio quality levels × three video quality levels).

To ensure consistent and correct audiovisual playback on a variety of commonly available computers, the 60 PVSs were very lightly compressed into Windows Media® Video format (WMV). The bitrate of these recompressed WMV files was selected independently of the PVS encoding bitrate so that additional artifacts would not be introduced (visible or audible). The audiovisual sequences available via CDVL are those compressed WMV sequences shown to subjects.

Subjects were asked to watch/listen to a series of audiovisual sequences and rate the overall audiovisual quality of each sequence. Ratings were collected using the absolute category rating (ACR) scale. ACR was implemented as a five-point, discrete quality rating scale: Excellent, Good, Fair, Poor and Bad.

Ten sets of audiovisual MOS values are available for vqegMM2 (see Table II). All six laboratories conducted the experiment in a controlled environment. Four labs repeated the experiment in a public environment. These opinion scores are available on CDVL.

Table II. Subjective Data Collection Summary

#	Lab	Environment	Total Subjects
1	NTIA	Controlled	28
2	NTIA	Public	9
3	Intel	Controlled	34
4	IRCCyN	Controlled	25
5	IRCCyN	Public	25
6	Technicolor	Controlled	24
7	Technicolor	Public	24
8	AGH	Controlled	14
9	AGH	Public	15
10	Opticom	Controlled	15

3. COMPONENT ANALYSES

Audio-only and video-only subjective ratings are not available for vqegMM2. Three analyses will be used to investigate these factors. These analyses use Difference Mean Opinion Score (DMOS) calculated with hidden reference removal, pooling all subjects from all labs. First, Table III shows the number of PVSs encoded at each combination of video bitrate and audio bitrate, in kbps.

Table III. Distribution of PVSs by Bitrate

Audio	Video Bitrate					
Bitrate	100 k	192 k	250 k	448 k	500 k	1000 k
64 k		6	1	4		6
32 k	1	5	1	5	1	4
8 k		5	1	4		6

Second, Table IV summarizes an ANOVA of the DMOS using the factors scene, video bitrate, and audio bitrate. Mean square error (MSE) measures the importance of each factor, where larger values indicate a greater impact on DMOS. Audio bitrate and video bitrate are the two most important variables and have similar importance.

Table IV. MSE for the Variables Scene, Video Bitrate, and Audio Bitrate

Scene	Video Bitrate	Audio Bitrate	Error	R^2
0.617	6.196	6.324	0.118	0.914

Third, objective quality models can be used to explain the vqegMM2 dataset. However, objective audio-only or video-only scores cannot fully characterize subjective audiovisual scores.

Pinson *et al* [2] indicates that a linearly fitted product

$$\hat{y} = \alpha + \mu\,(a \cdot v) \qquad (1)$$

provides an accurate prediction of audiovisual quality, where a is the audio MOS, and v is the video MOS. Constants α and μ depend upon the subjective scale. This theory draws upon 13 subjective tests performed independently by seven labs. The ANOVA of vqegMM2 indicates that Equation (1) captures most of the subjective information in the vqegMM2 dataset.

Table V identifies several objective audio quality or video quality models. Each model measures one aspect of quality by comparing the original and processed audiovisual sequences. Note that the models PEAQ (intended for high quality music) and POLQA (intended for speech only) are being applied outside of their intended scope. However, these are the most appropriate audio quality models available today for all audio content (e.g., music, speech, background noise, and sound effects). Table V also includes trivial models that use only the audio bitrate or video bitrate.

The objective models in Table V are separated into three categories: audio-only, video-only, and cross terms calculated using Equation (1). The model fit follows the VQEG High Definition Television (HDTV) Phase I test plan (see www.vqeg.org). The Pearson correlation is represented by ρ.

Table V. Performance of Objective Models

Audio Model	Contacts & Standards	ρ
PEAQ (DI)	ITU-T Rec. BS.1387, www.opticom.de	0.523
POLQA	ITU-T Rec. P.863, www.opticom.de	0.504
Audio Bitrate		0.489
Video Model	**Contacts & Standards**	**ρ**
PEVQ	ITU-T Rec. J.274, www.opticom.de	0.746
PSNR	ITU-T Rec. J.340, www.its.bldrdoc.gov	0.704
TetraVQM	Not validated, Barkowsky *et al.* [3]	0.705
VQM	ITU-T Rec. J.144, www.its.bldrdoc.gov	0.720
VQM_VFD	Not validated, www.its.bldrdoc.gov	0.780
Video Bitrate		0.696
Audio · Video		**ρ**
POLQA · PEVQ		0.855
PEAQ · PEVQ		0.855
POLQA · (1-VQM)		0.854
Audio · Video Bitrate		0.814

Table V demonstrates that video quality or audio quality alone is insufficient to describe audiovisual quality, yet a simple product will suffice. Higher order interactions are a topic for ongoing research. Model performance is suggestive but cannot be generalized, because vqegMM2 was not designed to stress models.

4. REFERENCES

[1] M. H. Pinson *et al.*, "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 6, No. 6, pp. 640–651, Oct. 2012.

[2] M. H. Pinson, W. Ingram and A. Webster, “Audiovisual quality components: an analysis,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 60-67, Nov. 2011.

[3] M. Barkowsky *et al.*, “Temporal trajectory aware video quality measure”. *IEEE Journal of Selected Topics in Signal Processing*, vol. 3 no 2, 2009, pp. 266–279.