An Objective Model for Audio-Visual Quality

Helard Becerra Martinez[b] and Mylène C.Q. Farias[a,b]

[a]Department of Electrical Engineering, University of Brasília (UnB), Brasília, Brazil;
[b]Department of Computer Science, University of Brasília (UnB), Brasília, Brazil;

ABSTRACT

In this paper, we describe three psychophysical experiments with the goal of understanding the influence of audio and video components on the overall perceived audio- visual quality. In Experiment I, subjects evaluated the quality of videos (without any audio) compressed at different video bitrates. In Experiment II, subjects evaluated the quality of audio (without any video) compressed at different audio bitrates. In Experiment III, subjects evaluated the quality of videos (audio- visual signals), which had their audio and video components compressed at different bitrates. The results of these experiments show that compressing only the video have a higher impact on the overall perceived quality than compressing only the audio. Another important goal of this paper is to propose an objective model for the audio-visual quality. Using the gathered data from Experiments I, II, and III, we are able to obtain two models with reasonably good correlations with the overall perceived quality.

Keywords: video quality assessment, audio and video quality, qos, qoe, video quality metrics.

1. INTRODUCTION

Digital video communication has evolved into an important field in the past few years. There have been significant advances in compression and transmission techniques, which have made possible to deliver high quality video to the end user. In particular, the advent of new technologies has allowed the creation of many new telecommunication services (e.g., direct broadcast satellite, digital television, high definition TV, Internet video). In these services, the level of acceptability and popularity of a given multimedia application is clearly related to the reliability of the service and the quality of the content provided. As a consequence, efficient real-time quality monitoring schemes that can faithfully describe the video experience — as perceived by the end user — is key for the success of these and future services.

The most accurate way to determine the quality of a video is by measuring it using psychophysical experiments with human subjects (subjective metrics).[1] Unfortunately, these experiments are expensive, time-consuming and hard to incorporate into a design process or an automatic quality of service control. Therefore, the ability to measure audio and video quality accurately and efficiently, without using human observers, is highly desirable in practical applications. With this in mind, fast algorithms that give a physical measure (objective metrics) of the quality are needed to obtain an estimate of the quality of a video when being transmitted, received or displayed.

Objective metrics represent a good alternative for measuring the video quality. This approach uses computational methods to process and evaluate the digital video and audio signal and to calculate a numeric value for the perceived quality. Unfortunately, within the signal processing community, quality measurements have been largely limited to a few objective measures, such as peak signal-to-noise ratio (PSNR) and total squared error (TSE). Although these metrics are relevant for data links and generic signals, in which every bit is equally important, they are not considered good estimates of the user's opinion about the received multimedia content because the outputs of these measures do not always correspond well with the human judgments of quality.[2,3]

Quality metrics that analyze visible differences between a test and a reference signal, taking into account aspects of the human visual system (HVS), usually have the best performance,[2,4] but are often computationally expensive and hardly applicable in real-time contexts. They can be classified according to the amount of reference (original) information used: Full Reference (FR), Reduced Reference (RR), and No-Reference (NR) metrics. On the FR approach the entire reference is available at the measurement point. On the RR approach, only part of the reference is available through an auxiliary channel. In this case, the information available at the measurement

point generally consists of a set of features extracted from the reference. Finally, on the NR approach the quality estimation is obtained only from the test video.

There is an ongoing effort to develop video quality metrics that are able to detect impairments and estimate their annoyance as perceived by human viewers.[5] To date, most of the achievements have been in the development of FR video quality metrics.[6–8] In particular, much remains to be done in the area of NR and RR quality metrics, which would certainly benefit from the incorporation of better perception models. In general, improvements can be achieved by incorporating better models for motion perception, pooling, and visual attention. A new trend in video quality design is the development of hybrid metrics, which are metrics that use a combination of the packet information, the bitstream header (without decoding), and the decoded video as inputs to the video quality estimation algorithm. With respect to applications, there is a great need for metrics that estimate the quality for multimedia applications. So far, very few metrics have addressed the issue of simultaneously measuring the quality of all media involved (e.g. video, audio, and text). Even for the simpler case of audio and video, there are only a few metrics in the literature that measure the quality of audio-visual content.[9–11]

We believe that in order to design good audio-visual metrics, it is necessary *first* to understand how audio and video contents are perceived. Most importantly, it is very important to understand how the degradations in audio and video affect the overall quality and how they interact with each other. Although over the years human responses to audio-video information have been studied extensively, research has been focused on determining the detection ability under different cross-modal presentation conditions.[12–15] For example, it has been shown that human sensitivity to audio-video asynchronies is not symmetrical.[12] Also, the presence of detectable audio-video temporal asynchronies results in a reduction of perceived quality.[13] Other findings showed that video quality influences subjective opinions of audio quality and vice-versa.[14,15]

The first goal of this paper is to obtain a better understanding of how audio and video components interact with each other and how these interactions affect the overall audio-visual quality. With this goal, we perform three psychophysical experiments and analyze their results. To generate the test sequences for these experiments, we start with original high definition video sequences with both audio and video components. For the first experiment, we consider only the video component of the sequences and compress them using a H.264 codec at different (video) bitrate values. For the second experiment, we consider only the audio component of the sequences and compress them using an MPEG-1 layer 3 codec, at different (audio) bitrate values. Finally, for the third experiment we consider both the sequence video and audio components and compress them independently. In all three experiments, we ask an average of 16 subjects to score the quality of the test sequences. Results of the experiments allow us to understand how the content of the videos affects the quality perceived by the final user.

The second goal of this work is to obtain an objective model for audio-visual quality. With this in mind, we test a set of combination models, using the scores of the first (only video) and second (only audio) experiments as basis to obtain the scores of the third (video and audio) experiment. To obtain the audio quality estimates, we use the no-reference audio quality metric SESQA (Single Ended Speech Quality Assessment Model).[16] To obtain the video quality estimates, we use a full-reference audio quality metric proposed by NTIA – The VQM (Video Quality Metric).[17] Then, we obtain two audio-visual quality metrics by combining these two metrics. The first using a linear model and the second using a Minkowski metric.

The paper is divided as follows. In Section 2, the psychophysical experiments are described. In Section **??**, the results of the experiments are discussed. In Section 3 the objective quality models for audiovisual quality are presented and discussed. Finally, in Section 4 the conclusions are presented.

2. SUBJECTIVE EXPERIMENTS

In this work, we performed three experiments. In *Experiment I*, subjects evaluated the quality of video signals (without any audio) compressed at different bitrates. In *Experiment II*, subjects evaluated the quality of audio signals (without any video) compressed at different bitrates. In *Experiment III*, subjects evaluated the quality of audio-video signals, which had their audio and video components independently compressed at different bitrates. In this section, we describe the apparatus and physical conditions, the content selection, the generation of test sequences, the experimental methodology, and the statistical methods used for the experiments performed in this work.

Table 1. Technical specifications of monitors and earphones used in the subjective experiments.

Monitor 1	Samsung SyncMaster P2370 Resolution: 1,920×1,080; Pixel-response rate: 2ms; Contrast ratio: 1,000:1; Brightness: 250cd/m2
Monitor 2	Samsung SyncMaster P2270 Resolution: 1,920×1,080; Pixel-response rate: 2ms; Contrast ratio: 1,000:1; Brightness: 250cd/m2
Earphones	Philips SHL580028 Headband Headphones Sensitivity: 106dB; Maximum power input: 50mW; Frequency response: 1028,000Hz; Speaker diameter: 40mm.

2.1 Apparatus and Physical Conditions

The experiments were run with two subjects at a time, using two separate PC desktop computers, two LCD monitors and two sets of earphones. The specifications of the monitors and earphones are shown in Table 1. The dynamic contrast of the monitors was turned off and the contrast was set at 100 and the brightness at 50. The room was sound proof and had the lights completely dimmed off to avoid any light to be reflected on the monitors.

The two desktop computers and their respective monitor screens were placed in a large table. Two chairs were placed in front of the two monitors. Two small lamps were placed close to the keyboards to help subjects enter their responses. The subjects were seated straight ahead of the monitor, centered at or slightly below eye height for most subjects. The distance between the subject's eyes and the video monitor was set at 3 screen heights. Three screen heights is a conservative estimate of the viewing distance according to the ITU-T Recommendation BT.500.[1] The software Presentation from Neurobehavioral Systems Inc. was used to run the experiment and record the subject's data.

Our subjects were voluntaries from the University of Brasília (UnB), Brazil. Most subjects were graduate students of the Departments of Computer Science and Electrical Engineering. They were considered naïve of most kinds of digital video defects and the associated terminology. No vision test was performed on the subjects, but they were asked to wear glasses or contact lenses if they needed them to watch TV. In each experiment, at least 15 subjects were used to guarantee robust results.

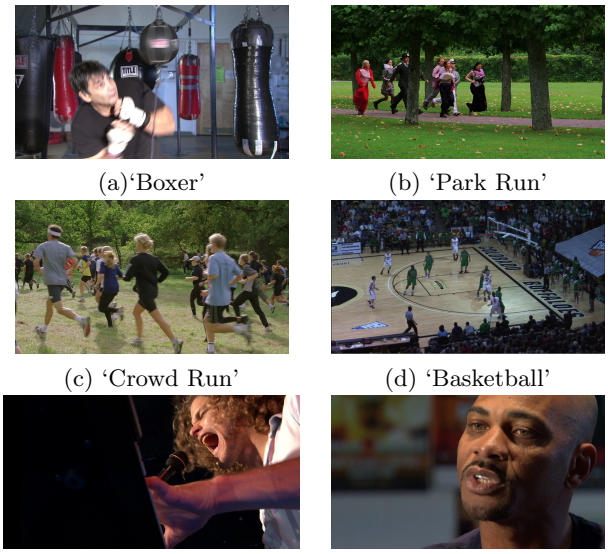
2.2 Content Selection

The original video sequences used in this work were obtained from The Consumer Digital Video Library (CDVL) website (`http://www.cdvl.org/`). The videos were high definition videos with spatial resolution of 1280x720, temporal resolution of 30 frames per second (fps), Color Sampling 4:2:0, and time duration equal to 8 seconds. All videos had accompanying audio. Nine video sequences were included in the experiments: 3 of them were used only in the trial and training sessions, while the other 6 videos were used in the main experimental sessions.

To choose the test sequences we followed the recommendations of The Final Report of VQEG on the validation of objective models multimedia quality assessment (Phase I), which states that the set of video sequences should have a good distribution of spatial and temporal activity.[18] We also took into account the audio content, selecting sequences that had speech, music and ambient sound. Representative frames of all 6 test sequences used in the main experimental sessions are presented in Figure 1.

Figure 2 shows the Spatial (SI) and Temporal (TI) perceptual information measures (computed as defined by Ostaszewska and Kloda[19]) for all original videos. As can be noticed in this figure, the video 'Reporter' has the highest temporal activity and the lowest spatial activity. The video 'Music' has both a high temporal activity and a high spatial activity, while the video 'Park Run' has relatively low spatial and temporal activities.

To give a description of the audio content of the original videos, we used the algorithm proposed by Giannakopoulos *et al.*[20] This algorithm divides the audio streams into several non-overlapping segments and, then, classifies each segment into one of the following classes: Music, Speech, Others1 (low environmental sounds: wind, rain etc), Others2 (sounds with abrupt changes, like a door closing), Others3 (louder sounds, mainly machines



(a)'Boxer' (b) 'Park Run'

(c) 'Crowd Run' (d) 'Basketball'

(e) 'Music' (f) 'Reporter'

Figure 1. Sample frames of original videos used in the subjective experiments.

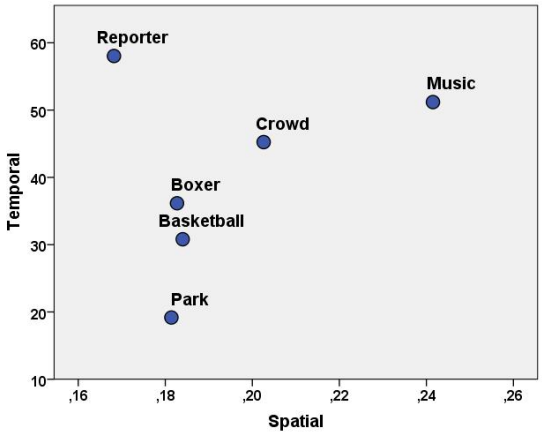


Figure 2. Spatial (SI) and Temporal (TI) perceptual information measures[19] of test sequences used in the subjective experiments.

and cars), Gunshots, Fights and Screams.[20] In Figure 3, the audio classification for the test sequences used in the experiments is presented, allowing a better comprehension of the audio components of each video sequence. As can be observed from the graph, the videos contain a good distribution of different audio types. The video 'Reporter' was classified mostly as Speech and partly as Others1. The video 'Park Run' was completely classified as Music, while the 'Music' video was classified as Others2, Music, and Screams. The videos 'Basketball' and 'Crowd Run' were both classified as Others1.

2.3 Generation of Test Sequences

The ffmpeg multimedia framework was used to encode all the video sequences used in the experiments. We have chosen video sequences of 8 seconds since this duration is an acceptable period of time for an observer to make a quality evaluation. Similar experiments have used sequences of 6 seconds with good results.[10]

For Experiment I, each of the original video test sequences (no audio) were compressed using the H.264 codec. Four different bitrate values were used: 30, 2, 1, and 0.8 Mbps. This test design resulted in 6 (original sequences) $\times$ 4 (bitrate values) = 24 test conditions. These 24 test conditions, plus 6 original sequences, resulted in the 30 test sequences shown during the main experimental session of Experiment I.

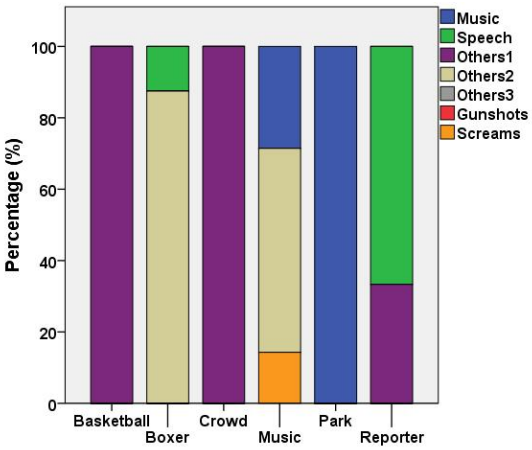


Figure 3. Audio classification of the test sequences used in the subjective experiments.

For the Experiment II, only the audio component of the videos was considered. The audio component was compressed using the MPEG-1 layer-3 coding standard. Three bitrate values were used: 128, 96, and 48 kbps. This test design resulted in 6 (original sequences) $\times$ 3 (bitrate values) = 18 test conditions. These 18 test conditions, plus 6 original audio signal, resulted in the 24 test sequences played during the main experimental session of Experiment II.

For Experiment III, both audio and video components of the test sequences were compressed. The video components were compressed with H.264, using the same bitrate values used in Experiment I. The audio components were compressed with MPEG-1 layer-3 coding standard, using the same bitrate values used in Experiment II. Considering the 3 bitrate values of the audio components and the 4 bitrate values of the video components (3 audio bitrates $\times$ 4 video bitrates) for all six sequences, resulted in a total of $3 \times 4 \times 6 = 72$ test conditions. These 72 test conditions, plus the 6 original sequences, completed the 78 test sequences shown during the main experimental session of Experiment III.

2.4 Experimental Methodology

An Absolute Category Rating With Hidden Reference (ACR-HR) methodology was used in all experiments.[1,21] Two sequences were presented in each trial and the source material was identical for both sequences. Of the two sequences, one was the reference and the other was the 'test' sequence. The subjects did not know which one was the reference and which one was the 'test' because the presentation order was randomized across trials. The subject was asked to give a quality score for each of the sequences in every trial.

Before starting the experiment, the experimenter made sure that the subject was properly seated at the adequate distance. The lamp was lit and the Presentation application was executed. For the experiment containing audio content (Experiments II and III), participants were asked to wear the earphones before the experiment began. A brief verbal explanation was made by the experimenter and the observer was asked to read the task instructions shown on the screen of the monitor.

The test was divided into three main sessions: Training, Practice, and Main sessions. In the *Training* session, the observer was shown a set of original sequences and the corresponding degraded sequences. The objective of this session was to familiarize the participant with the quality interval of the test sequences in the experiment. In the *Practice* session, the subject performed the same tasks performed in the Main session. The goal of the Practice session is to expose the subjects to sequences with a good range of impairments and to give subjects a chance to try out the data entry procedure. Since it takes time for a subject to get used to the task of judging/detecting impairments, the ITU Recommendation suggests that the first five to ten trials to be thrown away.[1] In our methodology, instead of discarding the first trials, we included 5 practice trials. Before beginning this stage, subjects were told that this is a Practice session and that no data is being recorded.

In the *Main* session, the actual task was performed. In the three experiments, observers were presented with a set of pairs of test conditions (audio, video, or audio-video). Upon completion of the presentation of both sequences, subjects were asked to rate them using a rating scale presented on the screen. They were told to disregard the content of the media and judge only the quality. The rating quality scale was between 0-100. In Figure 4, the continuous rating scale used in the experiments is shown. To avoid fatigue, the experimental session was broken in two parts.

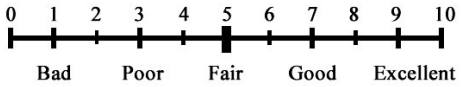


Figure 4. Rating scale used with DSCQS method. The pointer was located at the center of the scale.

2.5 Statistical Analysis Methods

The judgments given by the subjects to any test sequence are called subjective scores. This data is first processed by calculating the *mean observer score* (MOS) by averaging the scores over all observers for each test sequence:

$$\text{MOS} = \overline{S} = \frac{1}{L} \cdot \sum_{i=0}^{L} S(i), \tag{1}$$

where $S(i)$ is the score reported by the i-th subject, and L is the total number of subjects. For each test trial presented in the main experiment session, two quality scores were computed: one score for the test sequence and one score for the original sequence. We also calculated the sample standard deviation of the scores:

$$STD = \left(\frac{1}{L} \cdot \sum_{i=0}^{L} \left(S(i) - \overline{S} \right)^2 \right)^{1/2}, \tag{2}$$

and the internal standard error of $\overline{S}$:

$$\overline{STD} = \frac{STD}{\sqrt{L}}. \tag{3}$$

This is under the assumption that the scores are independent. The confidence interval for the 'true' MOS of a test sequence is given by $\overline{S} \pm \ t_{L,\alpha/2} \ \overline{STD}$, where $t_{L,\alpha/2}$ corresponds to the Student t coefficient.

As mentioned earlier, the videos in Experiment I had no audio and were compressed at different bitrates using an H.264 codec. In Experiment I, a total of 16 subjects scored the videos (without audio), generating one single MOS_v value for each test sequence. Figure 5 shows the obtained MOS_v versus the video bitrate (vb) values (vb1 = 800 Kbps, vb2 = 1 Mbps, vb3 = 2 Mbps, vb4 = 30 Mbps) for all test sequences.

As can be observed in this Figure 5, the MOS_v increases as the video bitrate increases. This shows that participants in this experiment were able to perceive variations in video bitrate (vb), which in turn resulted in variations in perceived video quality (MOS_v). The videos 'Basketball' and 'Park Run', which have both low temporal and spatial activities, showed the lowest MOS_v values, on average. The videos 'Music' and 'Crowd Run', which have both high temporal and spatial activities, got the highest MOS_v values, on average. These results are in agreement with the results in the literature that report that in more complex scenes (higher spatial and temporal activity) impairments are harder to see and, therefore, these scenes tend to be scored higher.

In Experiment II, the test sequences were formed of only audio components (no video). As described before, three audio bitrates (ab) were used. A total of 16 subjects scored the audio quality of the audio sequences in Experiment II, generating one MOS_a for each audio test sequence. Figure 6 shows the obtained MOS_a versus the audio bitrate values (ab1 = 48 kbps, ab2 = 96 kbps, ab3 = 128 kbps) for all test sequences. It can be seen that the MOS_a values increase as the audio bitrate values increase. The audio sequence 'Basketball', which was previously classified as Others1 (environmental sounds), presented the lowest MOS value. Meanwhile, the audio

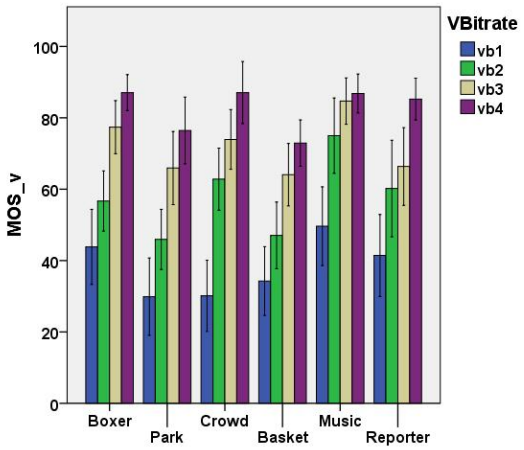


Figure 5. Experiment I: Mean Observer Values for Video (MOS_v) versus Bitrate, compressed video.

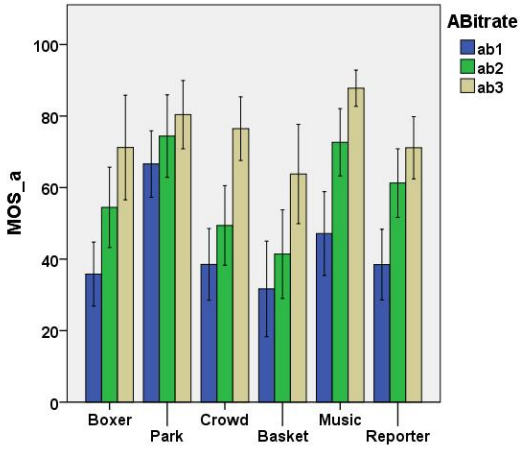


Figure 6. Experiment II: Mean Observer Values (MOS_a) versus Bitrate, compressed audio.

sequences 'Music' and 'Park Run' (classified as Music, Screams, and Others2) showed the highest MOS_a values. This seems to indicate that the audio quality of more complex sounds were less affected by audio compression.

In Experiment III, both audio and video components were included. Three audio bitrates and four video bitrates were used. A total of 17 subjects performed Experiment III, generating one MOS_{av} for each audio-visual test sequence. To understand the results of this experiment, we split the analysis in 2 parts.

First, Figure 7 shows how the MOS_{av} values change among all four video bitrate values, for different groups of 'originals' and audio bitrates. It can be observed that the MOS_{av} values increase as the video bitrate values increase, as in the two previous experiments. Nevertheless, the slope caused by the increase in video bitrate is *not* the same for the different 'originals' or the different groups of audio bitrates (ab). This can be observed for the sequences 'Boxer', 'Basketball' and 'Music', which have different slopes among different audio bitrates. Meanwhile, the sequences 'Park Run', 'Crowd Run', and 'Reporter' maintain similar slopes.

Second, Figure 8 shows that the MOS_{av} values change among all three audio bitrate values, for different groups of 'originals' and video bitrates. Again, it can be observed that the MOS_{av} values increase with the audio bitrate values. There are also differences in the behavior of the slope caused by the increase in audio bitrate. But, overall, the slopes of the increase are much smaller when compared to the slopes in Figure 7. In other words, compressing video had a higher impact on the overall quality than compressing audio.

Our last analysis consisted of trying to understand the contribution of the audio component to the overall

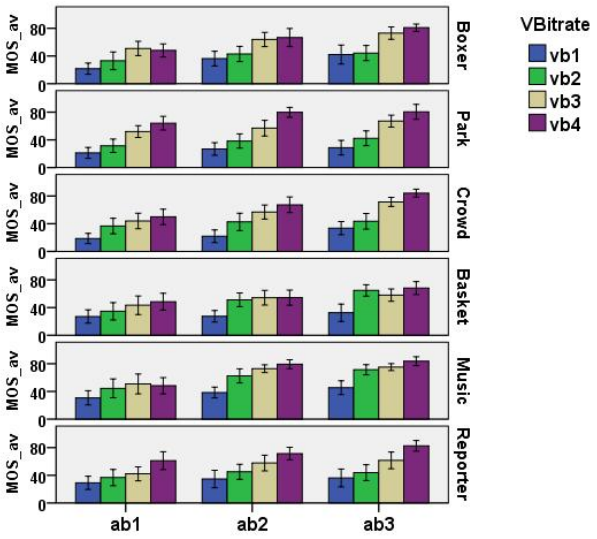


Figure 7. Experiment III: Mean Observer Values (MOS_{av}) versus audio bitrate.

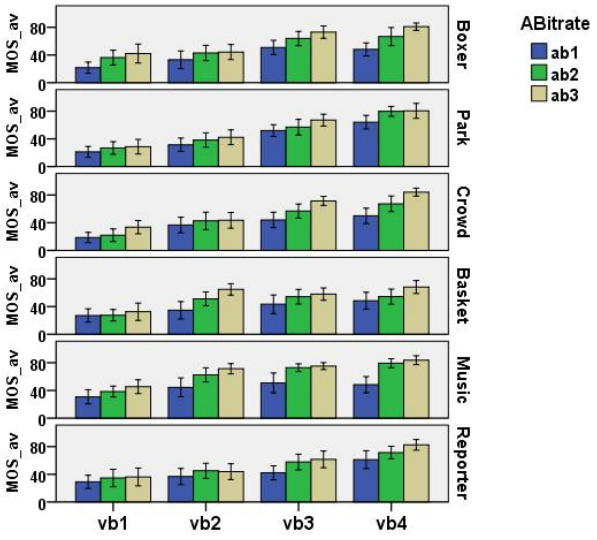
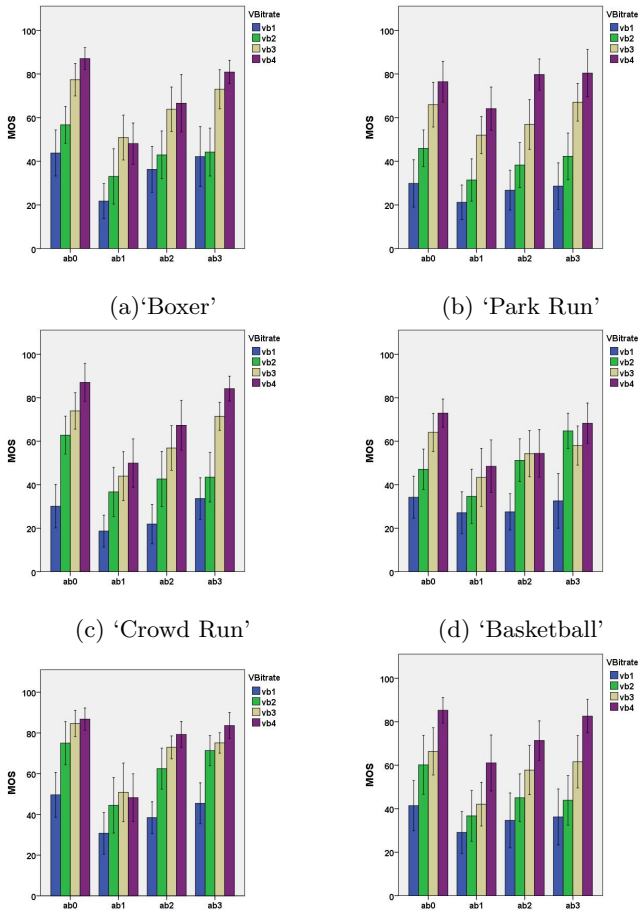


Figure 8. Experiment III: Mean Observer Values (MOS_{av}) versus audio bitrate.

quality. With this goal, we plotted the data from Experiment I and Experiment III in Figure 9. In this graph, the data from Experiment I (no audio) is shown as 'ab0' (first four columns in the left side of each graph). Notice from this figure that subjects rated the video sequences without any audio with *higher* MOS than the sequences with the audio – even high quality audio. Figure 9 shows that the absence of the audio component influences positively the perceived audio-visual quality. These results suggest that the audio component may be considered as a distraction factor for observers during a subjective test since, in this case, subjects have more things to pay attention to.

3. OBJECTIVE AUDIO-VISUAL QUALITY MODELS

To obtain the audio-visual quality model, we chose an audio quality metric and a video quality metric. The audio quality metric was the no-reference speech quality metric SESQA (Single Ended Speech Quality Assessment



(a)'Boxer' (b) 'Park Run'

(c) 'Crowd Run' (d) 'Basketball'

(e) 'Music' (f) 'Reporter'

Figure 9. Experiment I and III: MOS_v and MOS_{av} versus audio (and video) bitrates.

Model),[16] while the video quality metric was the full-reference metric VQM (Video Quality Metric).[17] In this section, we briefly describe the two objective models and the proposed objective audio-visual model.

3.1 Single Ended Speech Quality Assessment Model

The SESQA metric was originally proposed for speech signals in telephone applications. The first step of the SESQA algorithm consists of pre-processing the test signal, using a voice activity detector (VAD) that identifies speech signals and estimates its speech level. Then, the signal is analyzed and a set of 51 characteristic signal parameters is obtained. Next, based on a restricted set of key parameters, an assignment to main distortion classes is made. The main distortion classes include 'unnatural speech', 'noise', and 'interruptions, mutes, clippings'. The *key parameters* and the *assigned main distortion class* are used by the model to estimate the speech quality.

In order to apply this metric for audio signals (speech, music, generic sounds, etc.), we modified it slightly. Instead of using the 51 parameters considered in the original algorithm, we selected 17 parameters that showed better results for a set of degradade audio sequences. The content of these test sequences included music, explosion, speech, nature, etc. The rest of the original algorithm was kept without modifications.

3.2 Video Quality Metric

The Video Quality Metric (VQM) is a metric proposed by Wolf and Pinson from the National Telecommunications and Information Administration (NTIA) .[17] This metric has been adopted by ANSI as a standard for objective

video quality. In VQEG Phase II (VQEG, 2003), VQM presented a very good correlation with subjective scores, showing one of the best performances among the competitors.

The algorithm used by VQM includes measurements for the perceptual effects of several video impairments, such as blurring, jerky/unnatural motion, global noise, block distortion, and color distortion. These measurements are combined into a single metric that gives a prediction of the overall quality.

The VQM algorithm can be divided into the following stages:

- Calibration – Estimates and corrects the spatial and temporal shifts, as well as the contrast and brightness offsets of the processed video sequence with respect to the original video sequence.
- Extraction of quality features – The set of quality features that characterizes perceptual changes in the spatial, temporal, and chrominance domains are extracted from spatial-temporal sub-regions of the video sequen-ce. For this, a perceptual filter is applied to the video to enhance a particular type of property, such as edge information. Features are extracted from spatio-temporal (ST) subregions using a mathematical function and, then, a visibility threshold is applied to these features.
- Estimation of quality parameters – A set of quality parameters that describes the perceptual changes is calculated by comparing features extracted from the processed video with those extracted from the reference video.
- Quality estimation – The final step consists of calculating an overall quality metric using a linear combination of parameters calculated in previous stages.

3.3 Proposed Audio-Visual Model

In this section, we test a set of combination models using the *objective* scores for the video sequences of Experiment I and the *objective* scores for the audio sequences of Experiment II, as basis to obtain the predicted scores of Experiment III (video and audio). Both audio and video objective quality results were used as input to build the audiovisual quality model. With VQM and the adapted version of SESQA, we obtained objective scores for Experiments I and II, respectively. Both metrics showed reasonable correlation coefficients with the subjective scores of Experiments I and II: 0.82 and 0.94, respectively.

With the objective scores of Experiments I and II, we perform a regression analysis using the subjective data of Experiment III. The first model fitted was a simple linear model, given by the following equation:

$$\mathrm{Qav}_1 = \alpha_1 \cdot \mathrm{Qv} + \beta_1 \cdot \mathrm{Qa} + \gamma_1, \tag{4}$$

where Qav_1 corresponds to the predicted audiovisual quality score, Qv to the quality score obtained with VQM, and Qa to the quality score obtained with SESQA. The fit returned scaling coefficients $\alpha_1 = 0.45$, $\beta_1 = 0.48$, and $\gamma_1 = -8.9275$. For this fit, the Pearson correlation coefficient was 0.8472 and the Spearman correlation coefficient was 0.8337 (see Table 2). Figure 10 shows the graph of the predicted quality Qav_1 versus the subjective scores (MOS_{av}) for Experiment III.

The second model fitted to the data was the weighted Minkowski model, given by the following equation:

$$\mathrm{Qav}_2 = (\alpha_2 \cdot \mathrm{Qv}^p + \beta_2 \cdot \mathrm{Qa}^p)^{\frac{1}{p}}. \tag{5}$$

where Qav_2 corresponds to the predicted audiovisual quality score. Notice that if $p = 1$, this becomes the linear model with $\gamma_1 = 0$. The fit for the Minkowski model returned an exponent $p = 0.9165$ and scaling coefficients $\alpha_2 = 0.4184$ and $\beta_2 = 0.3999$. For this fit, the Pearson correlation coefficient was 0.8448 and the Spearman correlation coefficient was 0.8392 (see Table 2). Figure 11 shows the graph of the predicted quality Qav_2 versus subjective score (MOS_{av}) for Experiment III.

Finally, the third model fitted was a power model proposed by Wang and Bovik,[22] given the following equation:

$$\mathrm{Qav}_3 = (\gamma_2 + \alpha_3 \cdot \mathrm{Qv}^{p_1} \cdot \mathrm{Qa}^{p_2}), \tag{6}$$

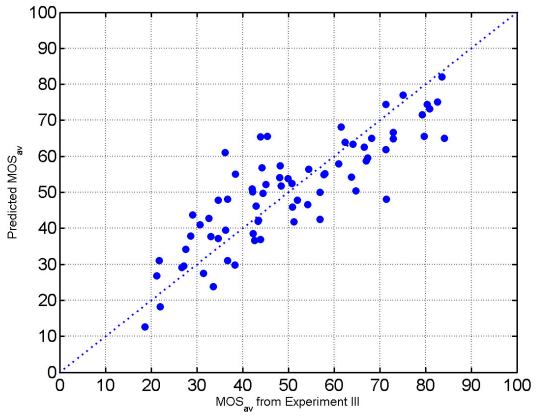


Figure 10. Predicted quality using linear model, Qav_1, versus subjective quality (MOS_{av}) for Experiment III.

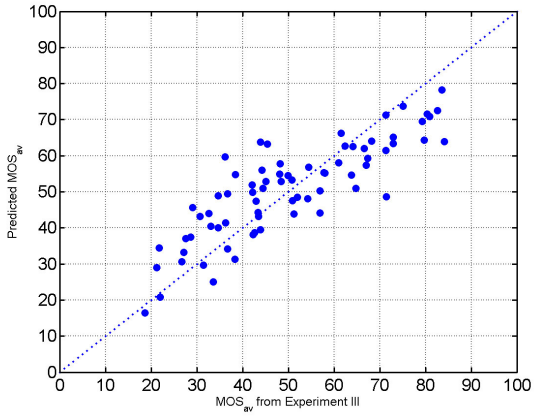


Figure 11. Predicted quality using the Minkowski model, Qav_2, versus subjective quality (MOS_{av}) for Experiment III.

where Qav_3 corresponds to the predicted audio-visual quality score. The fit for this model returned exponents $p_1 = 1.5837$ and $p_2 = 0.9524$ and scaling coefficients $\alpha_3 = 0.0006$ and $\gamma_2 = 26.9240$. For this fit, the Pearson correlation coefficient was 0.8545 and the Spearman correlation coefficient was 0.8384 (see Table 2). Figure 12 shows the graph of the predicted quality Qav_3 versus subjective quality (MOS_{av}) for Experiment III.

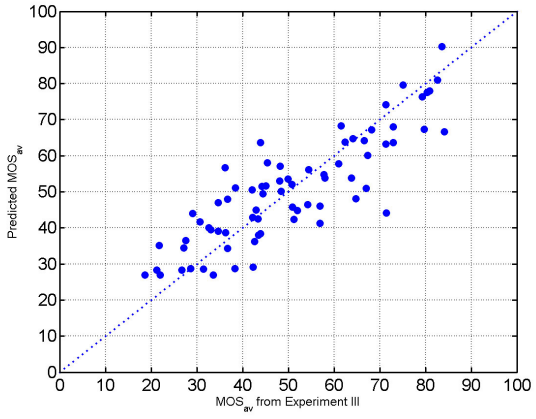


Figure 12. Predicted quality using the Power model, Qav_3, versus subjective quality (MOS_{av}) for Experiment III.

Table 2. Correlation Coefficients of audio-visual metrics tested on data of Experiment III.

Model	**Pearson**	**Spearman**
Qav_1	0.8472	0.8337
Qav_2	0.8448	0.8392
Qav_3	0.8545	0.8384
SQav_{H1}	0.8447	0.8340
SQav_{H2}	0.8441	0.8349
SQav_{G}	0.7739	0.8050
SQav_{W1}	0.8441	0.8349
SQav_{W2}	0.8244	0.8374

We can observe from the graphs that all models have a reasonnably good fit to the data. In Table 2, the Spearman and Pearson correlation coefficients of all models tested are listed, for the data of Experiment III. The differences among the different models is small and not consistent. For comparison purposes, we also tested the models by Hands,[10] Winkler and Faller,[23] and Garcia *et al.*[11] on the database of Experiment III. Hands proposed two *subjective* models, which are given by the following equations:

$$\text{SQav}_{\text{H1}} = 0.25 \cdot \text{MOS}_v + 0.15 \cdot (\text{MOS}_a \times \text{MOS}_v) + 0.95, \tag{7}$$

and

$$\text{SQav}_{\text{H2}} = 0.17 \cdot (\text{MOS}_a \times \text{MOS}_v) + 1.15, \tag{8}$$

where SQav_{H1} and SQav_{H2} are the predicted audio-visual quality score given by Hands' models.

Winkler and Faller[23] also proposed two *subjective* models given by the following equations:

$$\text{SQav}_{\text{W1}} = 0.103 \cdot (\text{MOS}_a \times \text{MOS}_v) + 1.98, \tag{9}$$

and

$$\text{SQav}_{\text{W2}} = 0.77 \cdot \text{MOS}_v + 0.456 \cdot \text{MOS}_a - 1.51, \tag{10}$$

where SQav_{W1} and SQav_{W2} are the predicted audio-visual quality score given by Winkler and Faller's models.

Finally, Garcia *et al.* proposed a subjective model give by the following equation:

$$\text{SQav}_{\text{G}} = 0.13 \cdot \text{MOS}_v + 0.0006 \cdot (\text{MOS}_a \times \text{MOS}_v) + 28.49, \tag{11}$$

where SQav_{G} is the predicted audio-visual quality score given by Garcia *at al.*'s model. These models are presented here only for comparison, given that they are all *subjective* models.

Observe from Table 2 that the proposed metrics (SQav_1, SQav_2, and SQav_3) have very similar correlation coefficients, which are all better than the other tested models (SQav_{H1}, SQav_{H2}, SQav_{W1}, SQav_{W2}, and SQav_{G}). The models SQav_{H1}, SQav_{H2} have the 4-th and 5-th best performance. It is worth pointing out these are subjective models, while the proposed models are objective models.

4. CONCLUSIONS AND FUTURE WORK

Three psychophysical experiments were conducted to understand the contribution of the audio and video components to the overall audio-visual perceptual quality. It was observed that the video content characteristics were important while determining the MOS, proving that there is a correlation between spatial and temporal activity and the MOS values gathered from experiments. By making an analysis of the audio content, we concluded that audio sequences classified as Others1 (low environmental sounds) were more sensitive to the compression than the other types of audio sequences. By observing the audio and video MOS results separately, it was possible to observe that the video compression had a higher impact on the overall audio-visual quality than the audio compression. Comparing the results of Experiments I and II, it was possible to see that the audio acted as a distractor factor, decreasing the MOS.

Using a video and an audio metrics, we were able to obtain three objective audio-visual quality models: a linear model, a weighted Minkowski model, and a power model. All models presented good fits with the subjective data, with correlation coefficients above 0.84. These objective models are very simple and can be used to predict the quality of audio-visual signals, given that we have an audio and a video quality metric. Further studies are needed in order to better understand how the content of the video and audio interact with each other and affect the audio-visual quality.

ACKNOWLEDGMENTS

The authors would like to thank all students of the Departments of Computer Science and Electrical Engineering that took part into the three Experiments. This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), in part by Universidade de Brasília (UnB), and in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

[1] ITU Recommendation BT.500-8, "Methodology for subjective assessment of the quality of television pictures," tech. rep. (1998).

[2] Moorthy, A. and Bovik, A., "Visual quality assessment algorithms: what does the future hold?," *Journal of Multimedia Tools & Applications* **51**, 675 – 696 (February 2011).

[3] Girod, B., "What's wrong with mean-squared error?," in [*Digital images and human vision*], 207 – 220, MIT Press (1993).

[4] Lin, W. and Kuo, C.-C. J., "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation* **22**(4), 297 – 312 (2011).

[5] Chikkerur, S., Sundaram, V., Reisslein, M., and Karam, L., "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Journal BC* **57**(2), 165–182 (2011).

[6] Daly, S., "The visible differences predictor: an algorithm for the assessment of image fidelity," in [*Digital Images and Human Vision*], Watson, A. B., ed., 179–206, MIT Press, Cambridge, Massachusetts (1993).

[7] Pinson, M. and Wolf, S., "An objective method for combining multiple subjective data sets," in [*Proc. SPIE Conference on Visual Communications and Image Processing*], **5150**, 583–92 (2003).

[8] Wang, Z., Lu, L., and Bovik, A., "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Comm.* **vol19**, 121–132 (2004).

[9] Soh, K. and Iah, S., "Subjectively assessing method for audiovisual quality using equivalent signal-to-noise ratio conversion," *Trans. Inst. Electron., Inform. Commun. Eng. A* **11**, 1305–1313 (2001).

[10] Hands, D. S., "A Basic Multimedia Quality Model," *Multimedia, IEEE Transactions on* **6**(6), 806–816 (2004).

[11] Garcia, M. N., Schleicher, R., and Raake, a., "Impairment-factor-based audiovisual quality model for iptv: Influence of video resolution, degradation type, and content type," *EURASIP Journal on Image and Video Processing* , 1–14 (2011).

[12] Bushara, K. O., Grafman, J., and Hallett, M., "Neural correlates of auditoryvisual stimulus onset asynchrony detection," *The Journal of Neuroscience* **21**(1), 300–304 (2001).

[13] Beerends, John G.; De Caluwe, F. E., "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc* **47**(5), 355–362 (1999).

[14] Steinmetz, R., "Human perception of jitter and media synchronization," *Selected Areas in Communications, IEEE Journal on* **14**(1), 61–72 (1996).

[15] Storms, R. L. and Zyda, M. J., "Interactions in perceived quality of auditory-visual displays.," *Presence: Teleoperators & Virtual Environments* **9**(6), 557 – 580 (2000).

[16] Malfait, L., Berger, J., and Kastner, M., "P.563 amp;8212;the itu-t standard for single-ended speech quality assessment," *Audio, Speech, and Language Processing, IEEE Transactions on* **14**(6), 1924–1934 (2006).

[17] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *Broadcasting, IEEE Transactions on* **50**(3), 312–322 (2004).

[18] VQEG, "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I," tech. rep. (2008).

[19] Ostaszewska, A. and Kloda, R., "Quantifying the amount of spatial and temporal information in video test sequences," *Recent Advances in Mechatronics, Springer* , 11–15 (2007).

[20] Giannakopoulos, T., Pikrakis, A., and Theodoridis, S., "A multi-class audio classification method with respect to violent content in movies using bayesian networks," in [*Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*], 90–93 (2007).

[21] ITU-R, "Recommendation P.911 : Subjective audiovisual quality assessment methods for multimedia applications," tech. rep. (1998).

[22] Wang, Z., Sheikh, H. R., and Bovik, A., "No-reference perceptual quality assessment of jpeg compressed images," in [*Proceedings, IEEE International Conference on*], **1**, I–477–I–480 (2002).

[23] Winkler, S. and Faller, C., "Perceived audiovisual quality of low-bitrate multimedia content," *Multimedia, IEEE Transactions on* **8**(5), 973–980 (2006).