Journal of Electronic Imaging

JElectronicImaging.org

Full-reference audio-visual video quality metric

Helard Becerra Martinez
Mylène C. Q. Farias

SPIE. IS&T imaging.org

Full-reference audio-visual video quality metric

Helard Becerra Martinez[a,*] **and Mylène C. Q. Farias**[a,b,*]
[a]University of Brasília, Department of Computer Science, Campus Universitário Darcy Ribeiro, 70919-970 Brasília, DF, Brazil
[b]University of Brasília, Department of Electrical Engineering, Campus Universitário Darcy Ribeiro, 70919-970 Brasília, DF, Brazil

Abstract. The goal of this work is to present a full-reference (FR) audio-visual quality metric. We performed three psychophysical experiments in order 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. 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. Based on the data gathered from these experiments, we obtain a set of subjective models for audio-visual quality. Inspired by these subjective models, we propose a set of FR audio-visual quality metrics composed of a combination of a video quality metric and an audio quality metric. The proposed metrics have good performance and present better results when compared to simple FR video quality metrics. *© 2014 SPIE and IS&T* [DOI: 10.1117/1.JEI.23.6.061108]

Keywords: video quality assessment; audio and video qualities; video quality metrics.

Paper 14180SSP received Apr. 1, 2014; revised manuscript received Jul. 23, 2014; accepted for publication Aug. 18, 2014; published online Sep. 10, 2014.

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 it 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).[1,2] 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 qualities accurately and efficiently, without using human observers, is highly desirable for 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 signals and to calculate a numerical value for the perceived quality. Quality metrics 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.[3] To date, most of the achievements have been in the development of FR video quality metrics.[4–6] 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. With respect to applications, there is a great need for metrics that estimate perceptual quality for multimedia applications. So far, few objective 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-visual content, achievements are limited and, currently, few objective metrics have been proposed.[2,7]

To design good audio-visual metrics, it is first necessary to understand how audio and video contents are perceived. Most importantly, it is necessary to understand how the degradations in audio and video affect the overall quality and how audio and video components interact with each other. Research in this area has been focused on determining the detection ability under different cross-modal presentation conditions.[8–11] For example, it has been shown that human sensitivity to audio–video asynchronies is not symmetrical.[8] Other works show that video quality influences subjective opinions of audio quality and vice versa.[10,11] Also, the

*Address all correspondence to: Helard Becerra Martinez, E-mail: hlrdbm03@gmail.com; Mylène C. Q. Farias, E-mail: mylene@unb.br

presence of detectable audio–video temporal asynchronies results in a reduction of perceived quality.[9]

As detailed by Pinson et al.[2] and You et al.,[7] several experiments in the literature have proposed audio-visual quality models that explore the relationship between audio and video qualities, measured separately, and the overall quality.[9,12–19] Results show that both video quality and audio quality are important to overall perceived quality, but their importance may differ for different multimedia applications.[7] For the majority of the audio-visual quality models, an inclusion of a cross term (audio quality × video quality) provides good results.[2] Most studies report that, in most applications, video quality is the dominant component of the overall quality.[9,13,15,16] Others report that video and audio are equally important in the overall audio-visual quality.[2,14] Audio quality seems to be more dominant than video in applications for which the audio signal conveys most of the information, like, e.g., video conferences and music clips.[7]

Although there are several perceptual audio-visual quality models available in the literature, the number of objective audio-visual quality metrics is much lower. One example is the work of Garcia et al.,[20] which presents both a subjective model and a parametric objective quality metric. The quality metric uses network packet-losses parameters to estimate quality and can only be used for transmission scenarios.

One of the goals 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. Both test sequences and subjective scores will be publicly available at the website of the Group of Digital Signal Processing of the University of Brasília.[21]

The second goal of this work is to obtain an FR audio-visual quality metric. Based on the data gathered from these experiments, we obtain a set of subjective models for audio-visual quality. With the help of these subjective models, we propose an FR audio-visual quality metric composed of the combination of a video quality metric and an audio quality metric. To obtain the audio quality estimates, we use the audio quality metric single ended speech quality assessment (SESQA) model.[22] To obtain the video quality estimates, we use the FR video quality metric proposed by National Telecommunications and Information Administration (NTIA)—The VQM.[23] Then, we obtain three FR audio-visual quality metrics by combining these two metrics using the same combination models used by the subjective models.

This paper is divided as follows. In Sec. 2, the psychophysical experiments are described. In Sec. 3, the experimental results are presented and discussed. In Sec. 4, a set of subjective models based on the experimental data is presented. In Sec. 5, the proposed FR audio-visual quality metrics are presented and their performance is discussed. Finally, in Sec. 6, the conclusions are presented.

2 Subjective Experiments

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 three experiments performed in this work.

2.1 *Apparatus and Physical Conditions*

The experiments were run with two subjects at a time, using two separate personal computer 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, the contrast was set at 100 and the brightness at 50. The room was sound proof and had the lights completely dimmed to avoid any light reflected on the monitors.

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 three screen heights, which is a conservative estimate of the viewing distance according to the ITU-T Recommendation BT.500.[1] The software Presentation from Neurobehavioral Systems Inc. (Berkeley, California) was used to run the experiment and record the subject's data.

Our subjects were volunteers from the University of Brasília, 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. Regarding the hearing acuity of participants, no test was conducted. However, participants did not report any hearing difficulties during the experimental session.

2.2 *Content Selection*

The original video sequences used in this work were obtained from the Consumer Digital Video Library.[24] The videos were 8-s long, had a resolution of 1280×720, a color space of $4:2:0$, and a frame rate of 30 frames per second (fps). All videos had accompanying audio. Nine video sequences were included in the experiments: three of them were used only in the trial and training sessions, while

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

Monitor 1	Samsung SyncMaster P2370 Resolution: 1920×1080; pixel-response rate: 2 ms; contrast ratio: 1000:1; brightness: 250 cd/m^2
Monitor 2	Samsung SyncMaster P2270 Resolution: 1920×1080; pixel-response rate: 2 ms; contrast ratio: 1000:1; brightness: 250 cd/m^2
Earphones	Philips SHL580028 headband headphones Sensitivity: 106 dB; maximum power input: 50 mW; frequency response: 1028 Hz; speaker diameter: 40 mm

Fig. 1 Sample frames of original videos used in the subjective experiments: (a) "Boxer," (b) "Park Run," (c) "Crowd Run," (d) "Basketball," (e) "Music," and (f) "Reporter."

the other six videos were used in the main experimental sessions.

To choose the test sequences, we followed the recommendations of the Final Report of Video Quality Experts Group (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 activities.[25] We also took into account the audio content, selecting sequences that had speech, music, and ambient sound. Representative frames of all six test sequences used in the main experimental sessions are presented in Fig. 1.

Figure 2(a) shows the spatial and temporal perceptual information measures (computed as defined by Ostaszewska and Kloda[26]) 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.

We used the algorithm proposed by Giannakopoulos et al.[27] to obtain a description of the audio content. This algorithm divides the audio streams into several nonoverlapping segments and 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, and cars), gunshots, fights, and screams.[27] In Fig. 2(b), the audio classification of the originals is presented. 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*

For experiment I, each of the original video test sequences (no audio) was compressed using the H.264 codec. Four different bitrate values were used: 30, 2, 1, and 0.8 Mbps. This test design resulted in $6(\text{original sequences}) \times 4(\text{bit rate values}) + 6\,\text{originals} = 30\,\text{test conditions}$.

For experiment II, only the audio component of the videos was considered. The audio component was compressed

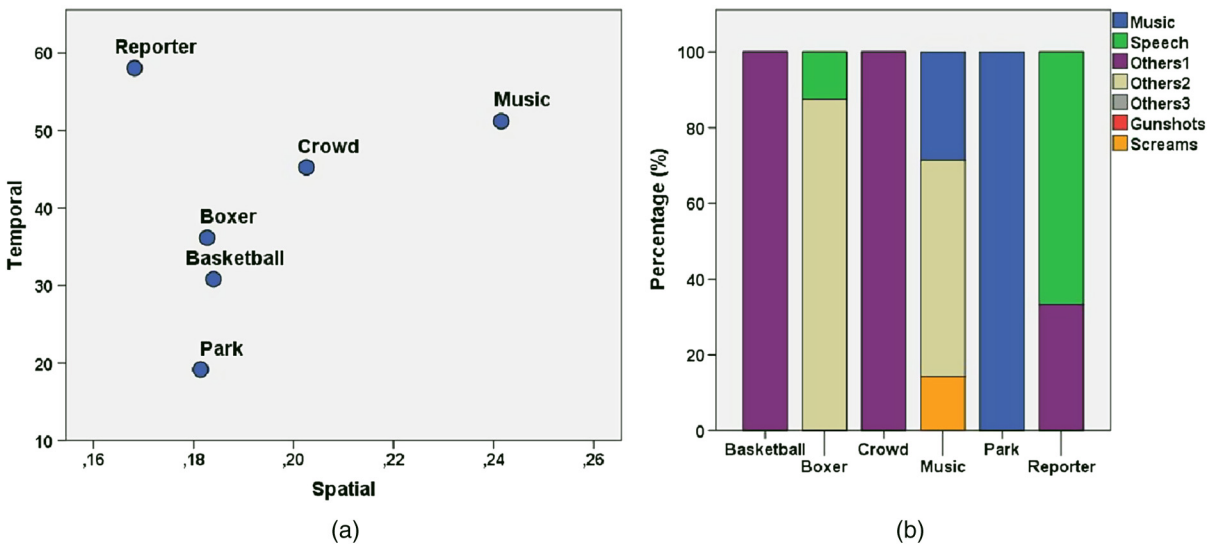

Fig. 2 (a) Spatial and temporal perceptual information measures[26] and (b) audio classification of the test sequences used in the subjective experiments.

using the MPEG-1 layer-3 coding standard. Three bitrate values were used: 128, 96, and 48 kbps. This test design resulted in $6(\text{original sequences}) \times 3(\text{bit rate values}) + 6\,\text{originals} = 24\,\text{test conditions}$.

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 (30, 2, 1, and 0.8 Mbps). The audio components were compressed with MPEG-1 layer-3 coding standard, using the same bitrate values used in experiment II (128, 96, and 48 kbps). Considering the three bitrate values of the audio components and the four bitrate values of the video components (3 audio bitrates × 4 video bitrates) for all six originals, this resulted in a total of $3 \times 4 \times 6 + 6\,\text{originals} = 78\,\text{test conditions}$.

2.4 *Experimental Methodology*

A double-stimulus continuous quality-scale methodology was used in all experiments.[1,28] Two sequences (with the same source material) were presented in each trial. Of the two sequences, one was the reference and the other was the "test" sequence. Subjects did not know which one was the reference and which one was the "test" because the presentation order was randomized across trials. After watching both sequences, subjects were asked to give a quality score for each of the sequences in every trial.

The test was divided into three main sessions: training, practice, and main sessions. In the training session, subjects were 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, subjects performed the same tasks performed in the main session. The goal of the practice session is to expose subjects to sequences with impairments and give them a chance to try out the data entry procedure. We included five practice trials.

In the main session, the actual task was performed. In the three experiments, after observers were presented with a set of pairs of test conditions (audio, video, or audio-video), they were asked to rate them using a quality scale between 0 and 100. The subject's participation time was limited to 30 min for experiment I, 25 min for experiment II, and 50 min for experiment III. A break was introduced in the middle of the main session to allow the subjects to rest.

2.5 *Statistical Analysis Methods*

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

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

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 the other score for the original sequence. We also calculated the sample standard deviation of the scores and the internal standard error of $\bar{S}$. When necessary, a t-test was performed to evaluate if differences in MOS were statistically significant.

3 Experimental Results

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 3(a) shows the obtained MOS_v versus the vb values (vb1 = 800 Kbps, vb2 = 1 Mbps, vb3 = 2 Mbps, vb4 = 30 Mbps) for all test sequences.

As can be observed in Fig. 3(a), MOS_v increases as the vb increases. This shows that participants in this experiment were able to perceive variations in vb, which in turn resulted in variations in perceived video quality (MOS_v). Considering the four levels of vb and the six different types of video contents (originals), we performed a univariate analysis of variance (ANOVA) on the video MOS_v. The analysis shows a main effect of the vb level ($F = 141.82$, $p < 0.01$) and of the video content ($F = 14.29$, $p < 0.01$). No interaction effect was detected between the factors "vb level" and "video sequence content." The videos "Basketball" and "Park Run," which have both low temporal and spatial activities, showed, on average, slightly lower MOS_v values (not statistically significant). The videos "Music" and "Crowd Run," which have both high temporal and spatial activities, got the highest MOS_v values on average. In these two scenes, some impairments might not have been perceived by the users due to the scene's characteristics and masking properties. In other words, errors of the same type and the same energy level (mean-squared error) when present in complex scenes have a higher visibility threshold than when present in lower activity scenes.[29]

In experiment II, the test sequences were formed of only audio components (no video). As described before, three audio bitrates (abs) 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 3(b) shows the obtained MOS_a versus the ab 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 ab values increase. Similar to what was done for experiment I, the same univariate ANOVA was computed for MOS_a. This analysis revealed a main effect of the ab level ($F = 63.93$, $p < 0.01$) and of the sequence type of content ($F = 13.56$, $p < 0.01$). No interaction effect was detected between the factors "ab level" and "audio sequence content." The audio sequence "Basketball," which was previously classified as others1 (environmental sounds), presented the lowest MOS value (not statistically significant). Meanwhile, the audio sequences "Music" and "Park Run" (classified as music, screams, and others2) showed the highest MOS_a values. This seems to indicate that degradations in more complex sounds are harder to perceive.

In experiment III, both audio and video components were included. Three abs and four vbs were used. A total of 17 subjects performed experiment III, generating one MOS_av for each audio-visual test sequence.

Figure 4(a) shows how the MOS_av values change among all four vb values for different groups of "originals" and abs. It can be observed that the MOS_av values increase as the vb values increase, as in the two previous experiments.

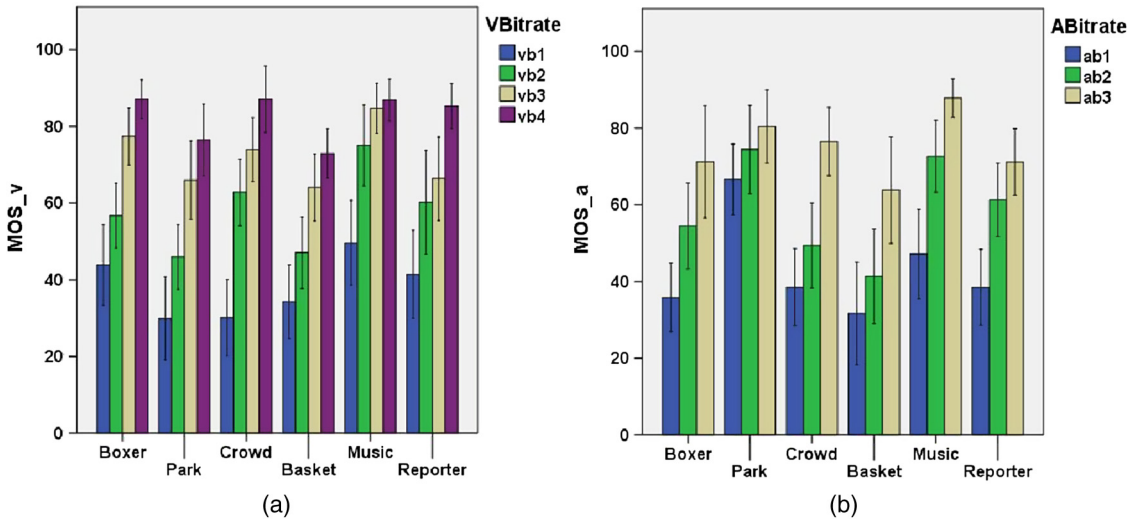

Fig. 3 (a) Experiment I: mean opinion values for video (MOS_v) versus bitrate, compressed video. (b) Experiment II: mean opinion values (MOS_a) versus bitrate, compressed audio.

Nevertheless, the slope caused by the increase in vb is not the same for the different "originals" or the different groups of abs. This can be observed for the sequences "Boxer," "Basketball," and "Music," which have different slopes among different abs. Meanwhile, the sequences "Park Run," "Crowd Run," and "Reporter" maintain similar slopes.

Figure 4(b) shows that the MOS_{av} values change among all three ab values for different groups of "originals" and vbs. Again, it can be observed that the MOS_{av} values increase with the ab values. There are also differences in the behavior of the slope caused by the increase in ab. But, overall, the slopes of the increase are much smaller when compared to the slopes in Fig. 4(a). 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 quality. With this goal, we plotted the data from experiment I and experiment III in Fig. 5. In these graphs, the data from experiment I (no audio) are shown as "ab0" (first four columns in the left side of each graph). Note that subjects rated video sequences without any audio with a slightly higher MOS value, especially for low audio quality sequences. In case of sequences with medium and high audio qualities, this difference is not statistically significant.

4 Subjective Quality Models

We used the subjective data gathered from experiments I, II, and III to obtain a set of three perceptual (subjective) models ($PrMOS_i$, $i = 1,2,3$) for the audio-visual quality (MOS_{av}), as a combination function of the audio quality (MOS_a) and the video quality (MOS_v).

The first subjective model tested was a simple linear model given by the following equation:

$$PrMOS_1 = \alpha_1 \cdot MOS_v + \beta_1 \cdot MOS_a + \gamma_1. \quad (2)$$

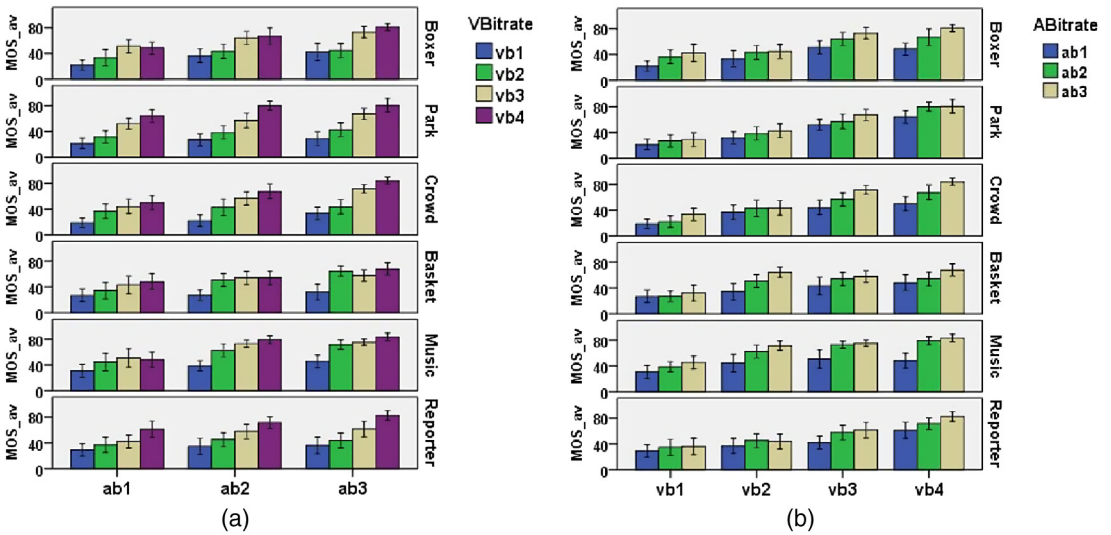

Fig. 4 Experiment III: (a) mean opinion values (MOS_{av}) versus audio bitrate (ab) and (b) mean opinion values (MOS_{av}) versus ab.

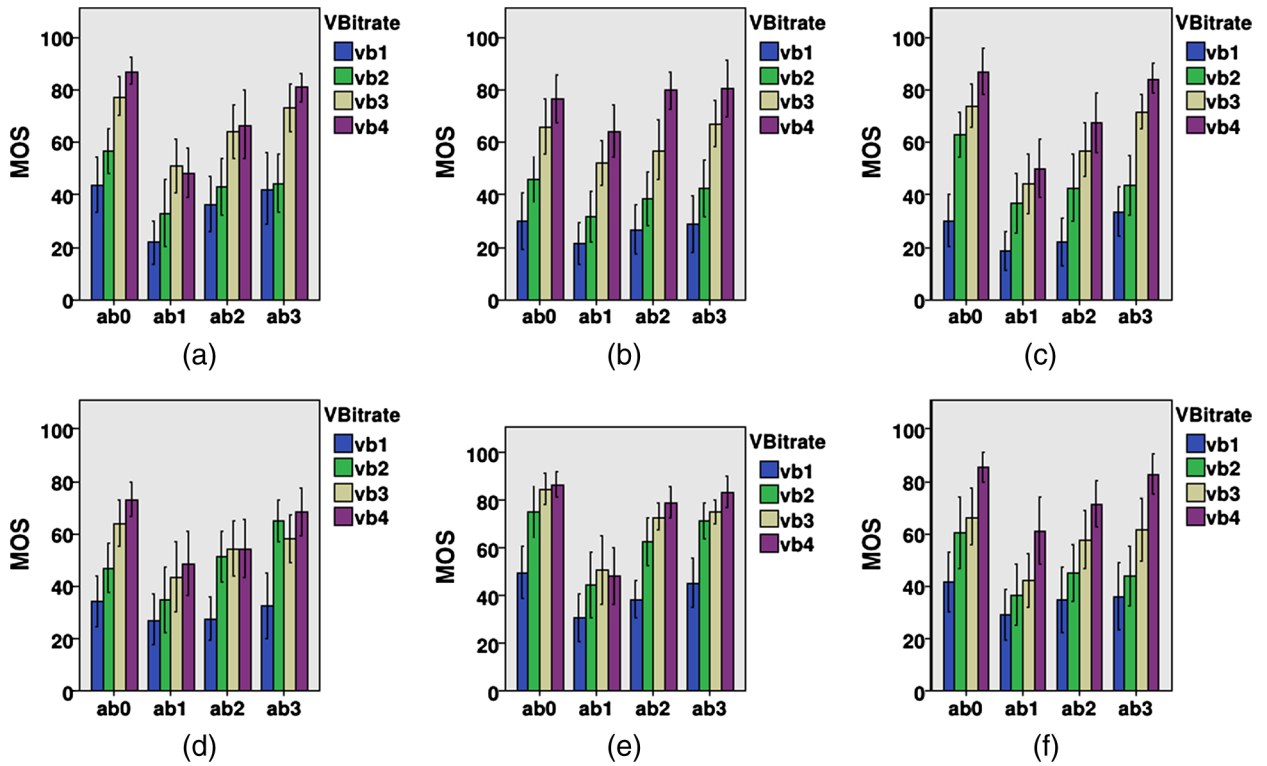

Fig. 5 Experiments I and III: MOS_v and MOS_{av} versus audio (and video) bitrates: (a) "Boxer," (b) "Park Run," (c) "Crowd Run," (d) "Basketball," (e) "Music," and (f) "Reporter."

The fitting returned scaling coefficients $\alpha_1 = 0.76$, $\beta_1 = 0.41$, and $\gamma_1 = -21.92$. The Pearson correlation coefficient (PCC) was 0.9110 and the Spearman correlation coefficient (SCC) was 0.9173.

The second model was a weighted Minkowski function given by the following equation:

$$\mathrm{PrMOS}_2 = (\alpha_2 \cdot \mathrm{MOS}_v^{p_1} + \beta_2 \cdot \mathrm{MOS}_a^{p_1})^{\frac{1}{p_1}}. \quad (3)$$

The fit returned $p_1 = 0.0001$, $\alpha_2 = 0.7024$, and $\beta_2 = 0.2976$. The PCC was 0.9197 and the SCC was 0.9267.

The last subjective model tested was a power model

$$\mathrm{PrMOS}_3 = (\gamma_2 + \alpha_3 \cdot \mathrm{MOS}_v^{p_2} \cdot \mathrm{MOS}_a^{p_3}). \quad (4)$$

The fit returned $p_2 = 1.3213$, $p_3 = 0.6533$, $\alpha_3 = -0.0109$, and $\gamma_2 = -12.9734$. The PCC was 0.9285 and the SCC was 0.9270.

We compared the subjective models obtained in this section with three subjective models available in the literature: two models ($\mathrm{SQav_{H1}}$ and $\mathrm{SQav_{H2}}$) proposed by Hands,[12] two models ($\mathrm{SQav_{W1}}$ and $\mathrm{SQav_{W1}}$) proposed by Winkler and Faller,[13] and one model ($\mathrm{SQav_G}$) proposed by Garcia et al.[20] Our goal here was to check which type of model has a good fit in our dataset. Our purpose was not to compare the models against each other. Given that these models were trained in different contents and different temporal and spatial resolutions, such a comparison would not be fair.

Hands' subjective model[12] was trained on sequences that include "head and shoulder" and "high-motion." The two subjective models proposed by Hands are given by the following equations:

$$\mathrm{SQav_{H1}} = 0.25 \cdot \mathrm{MOS_v} + 0.15 \cdot (\mathrm{MOS_a} \times \mathrm{MOS_v}) + 0.95 \quad (5)$$

and

$$\mathrm{SQav_{H2}} = 0.17 \cdot (\mathrm{MOS_a} \times \mathrm{MOS_v}) + 1.15, \quad (6)$$

where $\mathrm{SQav_{H1}}$ and $\mathrm{SQav_{H2}}$ are the predicted audio-visual quality scores.

Winkler's models[13] were trained on sequences destined for mobile applications which had very low audio and vbs. The two models are given by the following equations:

$$\mathrm{SQav_{W1}} = 0.103 \cdot (\mathrm{MOS_a} \times \mathrm{MOS_v}) + 1.98 \quad (7)$$

and

$$\mathrm{SQav_{W2}} = 0.77 \cdot \mathrm{MOS_v} + 0.456 \cdot \mathrm{MOS_a} - 1.51, \quad (8)$$

where $\mathrm{SQav_{W1}}$ and $\mathrm{SQav_{W2}}$ are the predicted audio-visual quality scores given by the models.

The model by Garcia et al.[20] was trained on high definition audio-visual sequences. This model is given by the following equation:

$$\mathrm{SQav_G} = 0.13 \cdot \mathrm{MOS_v} + 0.0006 \cdot (\mathrm{MOS_a} \times \mathrm{MOS_v}) + 28.49, \quad (9)$$

where $\mathrm{SQav_G}$ is the predicted audio-visual quality score given by Garcia's model.

Table 2 depicts the PCCs and SCCs obtained by testing all subjective models in the data of experiment III. As can be observed, the proposed power model ($\mathrm{PrMOS_3}$) presents the best results among all subjective models. The subjective

Table 2 Subjective audio-visual models: Pearson correlation coefficients (PCCs) and Spearman correlation coefficients (SCCs) obtained for data of experiment III.

Model	PCC	SCC
$PrMOS_1$	0.9110	0.9173
$PrMOS_2$	0.9197	0.9267
$PrMOS_3$	**0.9285**	**0.9270**
$PrMOS_{H1}$	0.8447	0.8340
$PrMOS_{H2}$	0.8441	0.8349
$PrMOS_G$	0.7739	0.8050
$PrMOS_{W1}$	0.8441	0.8349
$PrMOS_{W2}$	0.8244	0.8374

Note: Values in bold correspond to the models with best performance.

models taken from literature presented an acceptable correlation, given that they were not trained on this dataset.

To analyze how the three proposed subjective models ($PrMOS_1$, $PrMOS_2$, and $PrMOS_3$) perform for low and high quality contents, we classified the dataset according to their bitrates (audio and video). Video sequences were classified as low quality (vb1 and vb2 bitrates) and high quality (vb3 and vb4 bitrates). Audio sequences were classified as low quality (ab1 and ab2 bitrates) and high quality (ab3 bitrate). In Table 3, the PCCs for the different subsets of test sequences are presented, considering the same coefficients (parameters) used for the full set (results in Table 2). It can be observed that $PrMOS_2$ and $PrMOS_3$ perform better for low quality test material (lowest values of vb and ab). The worst PCC values were obtained for the highest values of ab and lowest values of vb.

A second analysis is presented in Table 4. In this case, we classified the dataset according to the video quality (MOS_a) and audio quality (MOS_v) values. Three quality classes were defined: low quality (0 to 33), medium quality (33 to 66), and high quality (66 to 100). It is observed that most of the sequences were rated above the medium quality class (over 33). Similarly, the highest values for the PCCs correspond to the sequences in the high quality class.

5 Audio-Visual Quality Metrics

To obtain the audio-visual quality metric, we combined an audio quality metric and a video quality metric. The audio quality metric was the speech quality metric SESQA,[22] while the video quality metric was the FR metric VQM.[23] In this section, we briefly describe the audio quality metric and video quality metric and the proposed objective audio-visual FR metric.

5.1 *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 preprocessing the test signal using a voice activity detector 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, assignment to the 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 in a test a set of degraded audio sequences. This set of audio sequences was different from the set used in the experiments and included sounds of music, explosion, speech, and nature. The set of 17 selected parameters is presented in Table 5. The rest of the SESQA algorithm was kept without modifications.

After training it, we tested SESQA using the audio signals of experiment II. Figure 6(a) shows the graph of MOS_a versus SESQA. The PCC is 0.9298 and the SCC is 0.9477. For comparison purposes, we also tested the performance of another audio metric: PEAQ.[30] Figure 6(b) shows the graph of MOS_a versus PEAQ. The PCC and SCC are both 0.4811. Therefore, SESQA perfoms better for the type of impairments considered in this work.

Table 3 PCCs of subjective models tested on low and high quality material subsets.

Video bitrate	Audio bitrate	Number of sequences	PCC $PrMOS_1$	PCC $PrMOS_2$	PCC $PrMOS_3$
Low (vb1, vb2)	All (ab1, ab2, ab3)	36	0.8050	0.8178	0.8214
	Low (ab1, ab2)	24	0.8227	0.8539	0.8540
	High (ab3)	12	0.6971	0.7268	0.7307
High (vb3, vb4)	All (ab1, ab2, ab3)	36	0.8602	0.8769	0.8944
	Low (ab1, ab2)	24	0.7891	0.8161	0.8441
	High (ab3)	12	0.9034	0.9119	0.8933

Table 4 PCCs of subjective models tested on different quality level subsets.

Video quality	Audio quality	Number of sequences	PCC $PrMOS_1$	PCC $PrMOS_2$	PCC $PrMOS_3$
Low (0 to 33)	All (0 to 100)	6	0.8261	0.8232	0.8288
	Low (0 to 33)	0	—	—	—
	Middle (33 to 66)	2	—	—	—
	High (66 to 100)	4	0.7900	0.8418	0.8309
Middle (33 to 66)	All (0 to 100)	33	0.7218	0.7313	0.7317
	Low (0 to 33)	3	—	—	—
	Middle (33 to 66)	17	0.6726	0.6517	0.6633
	High (66 to 100)	13	0.8471	0.8282	0.8447
High (66 to 100)	All (0 to 100)	33	**0.8602**	**0.8769**	**0.8944**
	Low (0 to 33)	1	—	—	—
	Middle (33 to 66)	17	0.6552	0.7032	0.7359
	High (66 to 100)	15	0.7580	0.7692	0.7533

Note: Values in bold correspond to the models with best performance.

5.2 *Video Quality Metric*

The video quality metric (VQM) is a metric proposed by Wolf and Pinson from the NTIA.[23] This metric has been adopted by American National Standards Institute 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 caused by 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.

5.3 *Proposed Audio-Visual FR Quality Metric*

We propose three FR audio-visual quality metrics, which are based on the subjective models described in Sec. 4. In other words, we use the same combination of models used as the subjective models to combine the audio and video metrics and predict the audio-visual quality. To obtain the coefficients, we use the subjective data of experiment III and the outputs of the audio quality metric and VQM.

The first model fitted was the simple linear model, given by the following equation:

$$\mathrm{Qav}_1 = \alpha_1 \cdot \mathrm{Qv} + \beta_1 \cdot \mathrm{Qa} + \gamma_1, \tag{10}$$

where Qav_1 corresponds to the predicted audio-visual 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 PCC was 0.8472 and the SCC was 0.8337 (see Table 6). Figure 7(a) shows the graph of the predicted quality Qav_1 versus the subjective scores ($\mathrm{MOS}_{\mathrm{av}}$) for experiment III.

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

$$\mathrm{Qav}_2 = (\alpha_2 \cdot \mathrm{Qv}^p + \beta_2 \cdot \mathrm{Qa}^p)^{\frac{1}{p}}, \tag{11}$$

where Qav_2 corresponds to the predicted audio-visual 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 PCC was 0.8448 and the SCC was 0.8392 (see Table 6). Figure 7(b) shows the graph of the predicted quality Qav_2 versus subjective score ($\mathrm{MOS}_{\mathrm{av}}$) for experiment III.

Finally, the third model fitted was a power model proposed by Wang et al.[31] given the following equation:

$$\mathrm{Qav}_3 = (\gamma_2 + \alpha_3 \cdot \mathrm{Qv}^{p_1} \cdot \mathrm{Qa}^{p_2}), \tag{12}$$

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 PCC was 0.8545 and the SCC was 0.8384 (see Table 6). Figure 7(c) shows the graph of the predicted quality Qav_3 versus subjective quality ($\mathrm{MOS}_{\mathrm{av}}$) for experiment III. We can observe from the graphs that all models have a reasonably good fit to the data.

Due to the difficulty of finding audio-visual quality metrics, we compared the proposed metrics with a group of FR video metrics. Although not completely fair, this comparison gives an idea of what performance gain can be obtained

Table 5 Selected 17 single ended speech quality assessment parameters (out of 51) used to calculate the audio quality.[22]

Parameter	Name	Classification
1	PitchAverage	Basic voice descriptors
2	SpeechLevel	Basic voice descriptors
3	MuteLength	Interruptions/mutes
4	LocalBGNoiseLog	Noise analysis
5	RelNoiseFloor	Noise analysis
6	SNR	Noise analysis
7	SpecLevelDev	Noise analysis
8	SpecLevelRange	Noise analysis
9	SpectralClarity	Noise analysis
10	BasicVoiceQuality	Unnatural speech
11	ArtAverage	Unnatural speech
12	CepCurt	Unnatural speech
13	FinalVtpAverage	Unnatural speech
14	LPCCurt	Unnatural speech
15	LPCSkew	Unnatural speech
16	PitchCrossCorrelOffset	Unnatural speech
17	PitchCrossPower	Unnatural speech

by also considering the audio quality, while estimating the audio-visual quality. The FR video quality metrics considered here are: structural similarity index (SSIM),[6] peak signal-to-noise ratio (PSNR), and video quality metric (VQM).[23] Although SSIM is a still-image quality metric, it has frequently been used for video. In fact, an implementation of SSIM is currently available in the H.264 codec. PSNR, on the other hand, is the most popular data metric and it is still in use by the signal processing community.

Table 6 PCCs and SCCs of FR audio-visual metrics tested on data of experiment III.

Model	PCC	SCC
Qav_1	0.8472	0.8337
Qav_2	0.8448	0.8392
Qav_3	**0.8545**	**0.8384**
SSIM	0.5896	0.6435
VQM	0.7092	0.7364
PSNR	0.5437	0.6350
$SQav_{H1}$	0.7707	0.7377
$SQav_{H2}$	0.7680	0.7371
$SQav_G$	0.7286	0.7809
$SQav_{W1}$	0.7682	0.7374
$SQav_{W2}$	0.7928	0.7973

Note: Values in bold correspond to the models with best performance.

In Table 6, the PCCs and the SCCs of all models are listed. For comparison purposes, the PCC values for the best subjective models are also presented. As can be observed, similarly to the subjective models, the proposed audio-visual quality metrics ($SQav_1$, $SQav_2$, and $SQav_3$) have the best correlation coefficients, with $SQav_3$ (power model) showing a slightly better result. Among the visual quality metrics (VQM, SSIM, and VQM), VQM presents the best correlation coefficients.

To analyze how the three proposed objective models ($SQav_1$, $SQav_2$, and $SQav_3$) perform for low and high quality contents, we classified the dataset according to their

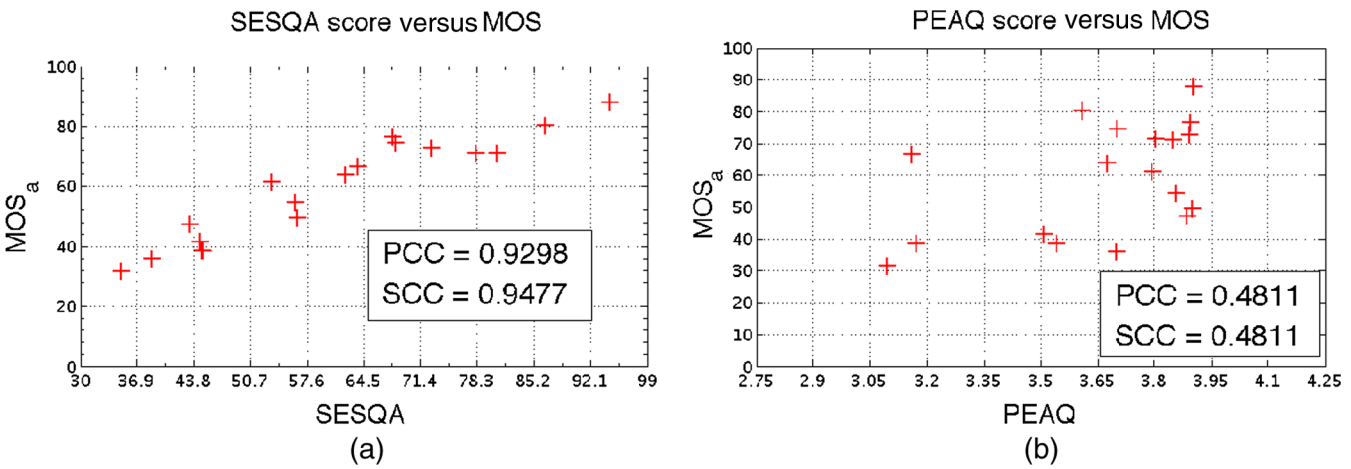

Fig. 6 Test of audio quality metrics: (a) MOS_a versus SESQA and (b) MOS_{av} versus PEAQ.

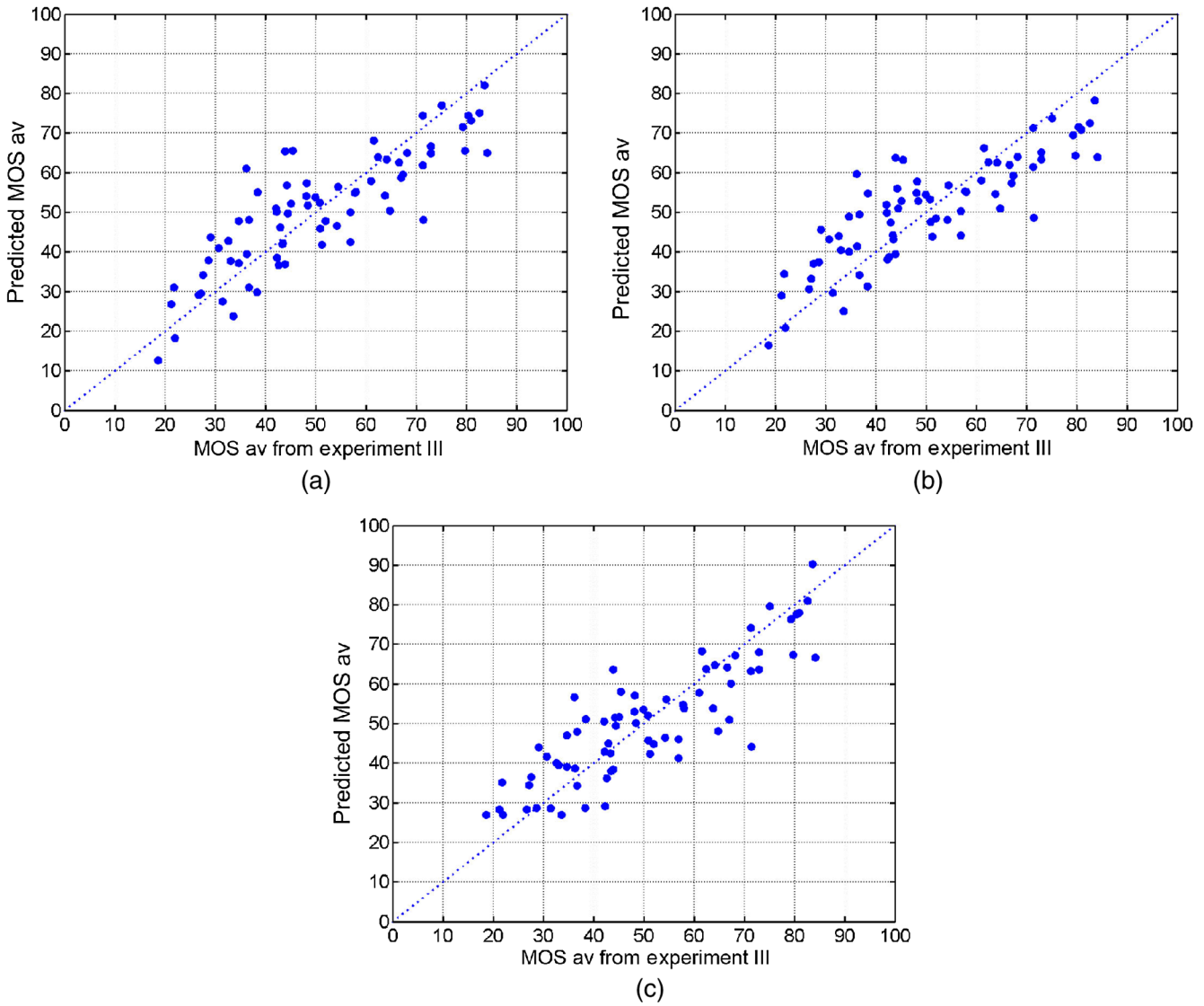

Fig. 7 Predicted quality using: linear model Qav_1, (b) Minkowski model Qav_2, and (c) power model Qav_3 for data of experiment III.

bitrates (audio and video). Video sequences were classified as low quality (vb1 and vb2 bitrates) and high quality (vb3 and vb4 bitrates). Audio sequences were classified as low quality (ab1 and ab2 bitrates) and high quality (ab3 bitrate). In Table 7, the PCCs for the different subsets of test sequences are presented, considering the same coefficients (parameters) used for the full set (results in Table 6). It can be observed that $SQav_1$ performs better with low quality test material (lowest vb and ab), while $SQav_3$ performs better for high quality material (highest vb and ab). The best PCC

Table 7 PCCs of FR audio-visual metrics tested on low and high quality material subsets.

Video bitrate	Audio bitrate	Number of sequences	PCC ($SQav_1$)	PCC ($SQav_2$)	PCC ($SQav_3$)
Low (vb1, vb2)	All (ab1, ab2, ab3)	36	0.7509	0.7445	0.7258
	Low (ab1, ab2)	24	0.7853	0.7761	0.7586
	High (ab3)	12	0.5881	0.5929	0.6404
High (vb3, vb4)	All (ab1, ab2, ab3)	36	**0.8276**	**0.8176**	0.8110
	Low (ab1, ab2)	24	0.8064	0.7927	**0.8187**
	High (ab3)	12	0.6181	0.6198	0.6486

Note: Values in bold correspond to the models with best performance.

Table 8 PCCs of FR audio-visual metrics tested on different quality level subsets.

Video quality	Audio quality	Number of sequences	PCC $PrMOS_1$	PCC $PrMOS_2$	PCC $PrMOS_3$
Low (0 to 33)	All (0 to 100)	9	0.6891	0.6777	0.5852
	Low (0 to 33)	0	—	—	—
	Middle (33 to 66)	4	—	—	—
	High (66 to 100)	5	—	—	—
Middle (33 to 66)	All (0 to 100)	24	0.5885	0.5873	0.5321
	Low (0 to 33)	0	—	—	—
	Middle (33 to 66)	16	0.7531	0.7485	0.7359
	High (66 to 100)	8	—	—	—
High (66 to 100)	All (0 to 100)	39	**0.8206**	**0.8282**	**0.8492**
	Low (0 to 33)	0	—	—	—
	Middle (33 to 66)	24	0.8162	0.8152	0.8250
	High (66 to 100)	15	0.5918	0.6216	0.6965

Note: Values in bold correspond to the models with best performance.

values were obtained for high values of vbs. On the other hand, the worst PCC values were obtained for the highest values of ab (both for high and low vbs). The model has a poor performance when the audio is high quality, with PCC values lower than those obtained for VQM. So, when the audio quality is high, a simple video quality metric performs better. When the audio quality is low, it has a bigger effect on the audio-visual quality.

As for the subjective quality models, an analysis considering the quality level results for the audio and video components is presented in Table 8. It is observed that the same pattern is repeated for the three quality metrics. The majority of sequences were scored with middle and high quality values as in the subjective quality models.

6 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 activities 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 compression degradations than other types of audio sequences. By separately observing the audio and video MOS results, it was possible to observe that the compression of the video component had a higher impact on the overall audio-visual quality than the compression of the audio component. Using a video metric and an audio metric, 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 PCCs 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 quality metric 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. Several aspects of audio-visual perception need special attention. For instance, research on the audio-visual quality perception from a neuro-physiological point of view will help to understand how both the visual and the auditory sensory channels are perceptually combined. Another aspect is the study of the cross-modal interactions between the audio and the video components and its dependency on the experimental context and, especially, on the audio-visual content. The study of the impact of audio-visual synchronization errors (e.g., lip synchronization) on audio-visual quality also needs further work.

Current projects focused on the development of audio-visual quality metrics, such as the audio-visual high definition quality project conducted by the VQEG, will certainly contribute to this research by providing new audio-visual models, new audio-visual materials, and reliable subjective data from experiments.

Acknowledgments

The authors would like to thank all students of the Departments of Computer Science and Electrical Engineering which took part in the three experiments. This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), in part by Universidade

de Brasília, 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," ITU-R Rec. BT.500, Int. Telecomm. Union, Geneva, Switzerland (1998).
2. M. Pinson, W. Ingram, and A. Webster, "Audiovisual quality components," *IEEE Signal Process. Mag.* **28**(6), 60–67 (2011).
3. S. Chikkerur et al., "Objective video quality assessment methods: a classification, review, and performance comparison," *IEEE J. Broadcast.* **57**(2), 165–182 (2011).
4. S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 179–206, MIT Press, Cambridge, Massachusetts (1993).
5. M. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," *Proc. SPIE* **5150**, 583–592 (2003).
6. Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," *Image Commun. Signal Process.* **19**(2), 121–132 (2004).
7. J. You et al., "Perceptual-based quality assessment for audio-visual services: a survey," *Image Commun.* **25**(7), 482–501 (2010).
8. K. O. Bushara, J. Grafman, and M. Hallett, "Neural correlates of auditory visual stimulus onset asynchrony detection," *J. Neurosci.* **21**(1), 300–304 (2001).
9. J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc.* **47**(5), 355–362 (1999).
10. R. Steinmetz, "Human perception of jitter and media synchronization," *IEEE J. Sel. Areas Commun.* **14**(1), 61–72 (1996).
11. R. L. Storms and M. J. Zyda, "Interactions in perceived quality of auditory-visual displays," *Presence Teleoperators Virtual Environ.* **9**(6), 557–580 (2000).
12. D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia* **6**(6), 806–816 (2004).
13. S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia* **8**(5), 973–980 (2006).
14. N. Kitawaki, Y. Arayama, and T. Yamada, "Multimedia opinion model based on media interaction of audio-visual communications," in *Proc. of the 4th Int. Conf. on Measurement of Speech and Audio Quality in Networks (MESAQIN'05)*, Prague, Czech Republic, pp. 5–10 (2005).
15. C. Jones and D. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," in *1998 Sixth International Workshop on Quality of Service(IWQoS 98)*, pp. 196–203, IEEE, Napa, CA (1998).
16. T. Hayashi et al., "Multimedia quality integration function for videophone services," in *IEEE Global Telecommunications Conf. 2007 (GLOBECOM'07)*, pp. 2735–2739, IEEE, Washington (2007).
17. M. Ries et al., "Audiovisual quality estimation for mobile streaming services," in *2nd Int. Symposium on Wireless Communication Systems*, pp. 173–177, IEEE, Siena (2005).
18. T. C. Thang, J. W. Kang, and Y. M. Ro, "Graph-based perceptual quality model for audiovisual contents," in *IEEE Int. Conf. on Multimedia and Expo*, pp. 312–315, IEEE, Beijing (2007).
19. K. Soh and S. Iah, "Subjectively assessing method for audiovisual quality using equivalent signal-to-noise ratio conversion," *Trans. Inst. Electron. Inform. Commun. Eng. A* **84**(11), 1305–1313 (2001).
20. M. N. Garcia, R. Schleicher, and A. Raake, "Impairment-factor-based audiovisual quality model for IPTV: influence of video resolution, degradation type, and content type," *EURASIP J. Image Video Process.* **2011**, 1–14 (2011).
21. "GPDS - Grupo de Processamento Digital de Sinais," http://www.gpds.ene.unb.br/mylene/databases.html (27 August 2014).
22. L. Malfait, J. Berger, and M. Kastner, "P.563: the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio Speech Lang. Process.* **14**(6), 1924–1934 (2006).
23. M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.* **50**(3), 312–322 (2004).
24. "The Consumer Digital Video Library (CDVL)," http://www.cdvl.org/ (27 August 2014).
25. VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Multimedia Quality Assessment, Phase I (2000).
26. A. Ostaszewska and R. Kloda, "Quantifying the amount of spatial and temporal information in video test sequences," in *Recent Advances in Mechatronics*, pp. 11–15, Springer, Poland (2007).
27. T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A multi-class audio classification method with respect to violent content in movies using Bayesian networks," in *IEEE 9th Workshop on Multimedia Signal Processing, 2007 (MMSP 2007)*, Crete, pp. 90–93, IEEE (2007).
28. ITU-R, *Recommendation P. 911: Subjective Audiovisual Quality Assessment Methods for Multimedia Applications* (1998).
29. S. Wolf and M. H. Pinson, "Spatial-temporal distortion metric for in-service quality monitoring of any digital video system," *Proc. SPIE* **3845**, 266–277 (1999).
30. T. Thiede et al., "PEAQ—the ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.* **48**(1–2), 3–29 (2000).
31. Z. Wang, H. R. Sheikh, and A. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *IEEE Int. Conf. on Proc.*, New York, Vol. 1, pp. 477–480, IEEE (2002).

Helard Becerra Martinez is a PhD student in the Department of Computer Science of the University of Brasília. He received his BS degree in computer science from Universidad Nacional San Antonio Abad del Cusco (UNSAAC), Peru, in 2010, and his MSc degree in computer science from University of Brasília, Brazil, in 2013. He is a researcher of the Digital Processing Signals Group (GPDS) and his current research interests include audio-visual signals, quality metrics, and image processing.

Mylène C. Q. Farias received her BSc in electrical engineering from the Universidade Federal de Pernambuco, Brazil, in 1995, her MSc in electrical engineering from the Universidade Estadual de Campinas, Brazil, in 1998, and her PhD in electrical engineering from the University of California Santa Barbara, USA, in 2004. She worked at CPqD (Brazil), Philips Research Laboratories (The Netherlands), and Intel Corporation (Phoenix, USA). Currently, she is a professor of electrical engineering at the University of Brasília.