Multimed Tools Appl
https://doi.org/10.1007/s11042-018-5656-7

Combining audio and video metrics to assess audio-visual quality

Helard A. Becerra Martinez[1] · Mylène C. Q. Farias[2]

Received: 13 March 2017 / Revised: 29 November 2017 / Accepted: 14 January 2018

Abstract In this work, we studied the use of combination models to integrate audio and video quality estimates to predict the overall audio-visual quality. More specifically, an overall quality prediction for an audio-visual signal is obtained by combining the outputs of individual audio and video quality metrics with either a linear, a Minkowski, or a power function. A total of 7 different video quality metrics are considered, from which 3 are Full-Reference and 4 are No-Reference. Similarly, a total of 4 audio quality metrics are tested, 2 of which are Full-Reference and 2 are No-Reference. In total, we tested 18 Full-Reference audio-visual combination metrics and 24 No-Reference audio-visual combination metrics. The performance of all combination metrics are tested on two different audio-visual databases. Therefore, besides analysing the performance of a set of individual audio and video quality metrics, we analyzed the performance of the models that combine these audio and video quality metrics. This work gives an important contribution to the area of audio-visual quality assessment, since previous works either tested combination models only on subjective quality scores or used linear models to combine the outputs of a limited number of audio and video quality metrics.

Keywords Video quality metrics · Audio quality metrics · Audio-visual quality metrics · Qoe · Multimedia quality assessment

✉ Helard A. Becerra Martinez
helardb@unb.br

Mylène C. Q. Farias
mylene@ieee.org

1 Department of Computer Science, University of Brasília (UnB), Campus Universitário Darcy Ribeiro, 70919-970 Brasília, DF, Brazil

2 Department of Electrical Engineering, University of Brasília (UnB), Campus Universitário Darcy Ribeiro, 70919-970 Brasília, DF, Brazil

Published online: 07 February 2018

1 Introduction

The great progress achieved by communications in the last twenty years is reflected by the amount of multimedia services available nowadays. One of the most popular multimedia services is the internet-based streaming, which has more recently gained an even bigger popularity. It is, nevertheless, understood that the success of this service relies heavily on its trustworthiness and on the quality of the content provided. Under these circumstances, the development of efficient real-time quality monitoring tools, which can quantify the audio-visual experience of multimedia services (as perceived by the end user) can bring real benefits to Internet Service Providers (ISP) and broadcast companies.

Psychophysical experiments are considered the most precise method to estimate the quality of audio-visual signals [10]. Unfortunately, these experiments are often expensive in terms of time and resources. Therefore, fast algorithms (objective quality metrics) arise as a good alternative for automatically determining the quality of audio-visual signals, as perceived by the end user [2]. To obtain a numerical estimate for the perceived quality, objective quality metrics use computational methods to process and evaluate signals. Depending on the amount of reference (original signal) information required by their algorithms, objective quality metrics can be classified as Full-Reference (FR), Reduced Reference (RR), and No-Reference (NR) metrics. In the case of FR metrics, the entire reference is needed at the measurement point to obtain the quality estimation. For the RR metrics, only a part of the reference is needed, which can be made available at the measurement point through an auxiliary channel. Finally, for the NR metrics the quality estimation is obtained blindly, using only the test video.

There is an ongoing effort to develop video quality metrics that estimate quality as perceived by human viewers, but most of the achievements have been in the development of FR video quality metrics [2, 11, 18]. Much remains to be done in the area of no-reference (NR) quality metrics [18]. Also, very few objective metrics have addressed the issue of simultaneously measuring the quality of multimedia content (e.g. video, audio, and text), as pointed out by Pinson et al. [20]. For the simpler case of audio-visual content, a lot of work has been done on trying to understand audio-visual quality, what resulted in several subjective models [6, 22]. But, only a few works tackle the problem of developing audio-visual objective quality metrics [5, 28].

In this work, we investigate how to assess the quality of audio-visual signals using combinations of simple audio and video quality metrics. The audio and video metrics are combined using three models: Linear, Minkowski, and Power functions. The combination models were inspired in the analysis of data gathered from 3 psychophysical experiments in which audio and video quality scores were collected. Using these combination models, we propose a set of FR and NR audio-visual quality assessment methods. Each method is composed by a video quality metric, an audio quality metric, and a model that combines the audio and video (objective) predictions to provide an overall audio-visual quality estimate. A total of 7 different video quality metrics are considered, from which 3 are FR and 4 are NR. Similarly, a total of 4 audio quality metrics are considered, 2 of them are FR and 2 are NR. The performance of these audio-visual quality methods is tested and validated using two audio-visual quality databases.

Besides presenting a performance analysis of a set of audio and video quality metrics, the results presented in this work contribute to a better understanding of how audio and video objective quality scores can be combined to predict the overall audio-visual quality. Given the mature state-of-the-art of audio and video quality metrics, we believe this is an important step towards the design of accurate audio-visual quality metrics. In this work, we explore the use of three combination models (Minkowski, linear, and power models) for audio-visual quality assessment. We tested these models using a set of video and audio quality metrics (both NR and FR), validating them on two different quality databases. We believe this work is an important contribution to the area of audio-visual quality assessment, given that previous works either tested combination models only on subjective scores or used only linear models to combine the outputs of a limited number of audio and video quality metrics.

This paper is divided as follows. Section 2 presents a brief description of some combination techniques used in previous studies. In Section 3, we describe the three psychophysical experiments that are part of the UnB Audio-Visual Quality (UnB-AVQ) Database 1. In Section 4, we present the combination models used to merge the audio and video predictions. In Sections 5 and 6, the FR and NR audio-visual quality assessment methods are presented. A performance analysis of the two approaches is carried out using the database described in Section 3. In Section 7, both groups of FR and NR audio-visual metrics are tested using the NTIA audio-visual quality database. Finally, in Section 8, our conclusions are discussed.

2 Related work

Several audio and video quality metrics have been proposed in the past few years. Several of these metrics present good performance levels, in terms of complexity and accuracy [1], but they are only capable of estimating either audio or video quality, but not both. Among the different approaches used in the design of quality metrics, a few methods use different models to combine the contributions of the most common degradations (artifacts) to produce the overall quality. For instance, Farias designed a no-reference (NR) video quality metric in which the overall annoyance is predicted by combining the outputs of blurring, blocking, and noise strength metrics [4]. One of the combinations models used in this work was a weighted Minkowski model. Additionally, Wang and Bovik [24] developed an objective NR image quality metric, targeted at JPEG compressed images, which combines the outputs of a blocking and a blurring strength metrics to estimate the overall image quality. The outputs of these two metrics were combined using a non-linear power model.

Given the progress achieved in the area of audio and video quality assessment (independently) [3], the next step is the design of an audio-visual quality metric. Considering that audio-only or video-only quality metrics cannot estimate audio-visual quality [21], a recent research trend is the use of models that combine the outputs of audio and video metrics to estimate audio-visual quality [20]. The first audio-visual combination models were tested on subjective quality scores [6, 21, 27]. Although these works cannot be used in real multimedia applications, their results have helped understand how individual audio and video quality estimates can be combined to predict the overall audio-visual quality.

Currently, there are only a few audio-visual objective quality metrics available in the literature. Up to our knowledge, most of them are parametric metrics, i.e. metrics that estimate quality using the information available at the receiver, such as bitrate, frame rate, quantization index, motion vectors, and network information. Among the currently available audio-visual parametric metrics, we can cite the works of Garcia et al. [5] and Yamagishi and Gao [28]. The parametric model proposed by Yamagishi and Gao [28], standardized in ITU-T Recommendation P.1201, uses information extracted from packet headers and network. Garcia et al. [5] proposed a parametric metric that uses impairment factors, which are extracted from the bitstream or packet headers, to quantify the overall quality. Although parametric metrics are faster than pixel-based video quality metrics, they are dependent on the type of coding and transmission process, what makes them less generally applicable. In other words, they cannot predict the quality of 'offline' content, like, for example, content transcoded among different compression standards/bitrates or processed using specific signal processing techniques.

It is worth pointing out that, in previous works, one of the most popular combination models is the linear model [5, 6, 27]. This model has the advantage of being very simple, however it does not provide a good accuracy performance. In fact, studies have shown that better accuracy performance can be obtained when a power model, which inludes a multiplicative cross term (audio quality $\times$ video quality), is used to predict audio-visual quality [20]. In this work, besides testing linear and power models, we also tested Minkowki models.

3 Psychophysical experiments

To design better audio-visual metrics, we first need to understand how audio and video components interact with each other and how these components can be combined to produce the overall audio-visual quality. With this goal, in this work, we use data collected from human observers/listeners who participated in three psychophysical experiments. Using the subjective responses from all participants we were able to measure the audio, video, and audio-visual quality of compressed audio-visual signals.

The experiments are part of the UnB Audio-Visual Quality (UnB-AVQ) Database 1. In these experiments, six original high definition video sequences (with audio and video components) from The Consumer Digital Video Library[1] are used. Representative frames of the original sequences are shown in Fig. 1. Each sequence is eight seconds long and has a spatial and temporal resolution of 1280×720 (720p) and 30 frames per second (fps) respectively. Each source sequence was compressed using four video bitrates and three audio bitrates. The video and audio components were individually compressed and, then, combined. The bitrate values were chosen to provide similar ranges of quality for the audio and video sequences. Specifications of the codecs, bitrates, and number of sequences are listed in Table 1. Detailed information regarding the UnB-AVQ database can be found in a previous work [14].

All three experiments were conducted following the International Telecommunications Union (ITU) recommendation ITU-R. BT-500 [10], which details the necessary equipment, the physical conditions, the selection of participants, and the experimental methodology. The experiments were run with two participants at a time. Therefore, two separate desktop computers, two LCD monitors, and two sets of earphones were installed in the room. Detailed specifications of the equipment used in the experiments are depicted in Table 2. Experiments took place in a recording studio (sound proof) with the lights completely dimmed to avoid any light reflection on the monitors. Distance between subject's eyes and the monitor was set at three screen heights (3H), in accordance with ITU-R. BT-500 [10].

[1] http://www.cdvl.org.

Fig. 1 Sample frames of the original videos from the UnB Audio-Visual Quality (UnB-AVQ) Database 1, available at http://www.ene.unb.br/mylene/databases.html

Participants were volunteers from the University of Brasilia, Brazil. They were mostly graduate students from the departments of Computer Science and Electrical Engineering. No particular vision or hearing test was performed on the participants. But, they were

Table 1 Detailed Specifications for Experiments I-III of UnB Audio-Visual Quality (UnB-AVQ) Database 1

Component	Experiment I Video	Experiment II Audio	Experiment III Audio + Video
Bitrate	30, 2, 1, 0.8 MB/s	128, 96, 48 KB/s	128, 96, 48 KB/s 30, 2, 1, 0.8 MB/s
Codec	H.264	MPEG-1 Layer 3	MPEG-1 Layer 3 H.264
# Test seq.	30	24	78
# Subjects	16	16	17

Download from: http://www.ene.unb.br/mylene/databases.html

Table 2 Technical specifications of monitors and earphones used in the subjective experiments

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

asked to wear glasses or contact lenses if they needed them to watch TV. The number of participants for each experiment is depicted in Table 1

Regarding the assessment method, a double-stimulus continuous quality-scale methodology was applied (ITU Recommendation BT-500 [10]). Such metho-dology implies that, for each trial of the experiment, two sequences (with the same content) are presented to the participant: a reference sequence and a test sequence. After these two sequences are presented (in random order), participants are asked to give a quality score for each sequence. Additionally, to familiarize the participant with the test procedure and guarantee reliable results, Display and Training sessions were included at the beginning of the experiment.

In Experiment I, subjects evaluated the quality of video (only) sequences compressed using the H.264 codec. In Experiment II, subjects evaluated the quality of audio (only) sequences compressed with MPEG-I layer-3 codec. Finally, in Experiment III, both audio and video components were independently compressed and subjects evaluated the overall audio-visual quality.

For all experiments, the quality scores were averaged over the subjects to produce a Mean Opinion Score (MOS) for each test sequence, presented in a 0 – 100 range. Figure 2 presents a scatter plot with results from the subjective experiments for each single component (audio and video). After analysing the experimental results, we observed that the bitrate of the video component has a higher impact on the global audio-visual quality than the bitrate of the audio component. Also, the characteristics of both video and audio content affect the

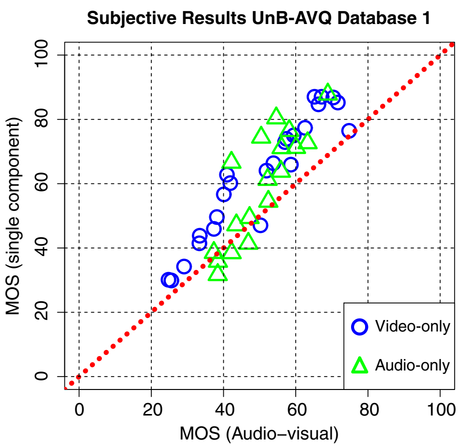

Fig. 2 Subective Results Unb AVQ Database 1, available at http://www.ene.unb.br/mylene/databases.html

perceived audio-visual quality [13, 14]. The videos and the corresponding subjective data of the UnB Audio-Visual Quality (UnB-AVQ) Database 1 are available for download at the website of the Group of Digital Signal Processing of the University of Brasilia.[2]

4 Perceptual quality models

Based on the results gathered from Experiments I-III, we developed a set of *subjective* audio-visual quality models. Similarly to what is found in the literature [6], three functions were used to combine the audio and video MOS values, referred as MOS_a and MOS_v), respectively.

The first subjective audio-visual quality model is a simple linear model, given by the following equation:

$$\mathrm{PrMOS}_1 = \alpha \cdot \mathrm{MOS}_v + \beta \cdot \mathrm{MOS}_a + \gamma. \tag{1}$$

The fitting returned three scaling coefficients denoted by α, β (video and audio regression coefficients, respectively), and γ (an intercept).

The second model is a weighted Minkowski function, given by:

$$\mathrm{PrMOS}_2 = (\alpha \cdot \mathrm{MOS}_v^p + \beta \cdot \mathrm{MOS}_a^p)^{\frac{1}{p}}. \tag{2}$$

Similarly, the fitting for the second model returned three coefficients denoted by α, β, (weight coefficients for video and audio, respectively) and ρ (a power coefficient).

The third subjective model is a power model, given by:

$$\mathrm{PrMOS}_3 = (\gamma + \alpha \cdot \mathrm{MOS}_v^p \cdot \mathrm{MOS}_a^p). \tag{3}$$

The fitting for the third model resulted in four coefficients, denoted by γ (an intercept coefficient), α (a weight coefficient), and ρ_1, ρ_2 (power coefficients for video and audio, respectively).

Pearson Correlation Coefficients (PCC) for all three perceptual models are depicted in Table 3. By comparing all three models results, we noticed that the power model (PrMOS_3) had a slightly better performance in terms of correlation, reaching a Pearson Correlation Coefficient (PCC) of 0.92. Further analysis showed that the models PrMOS_2 and PrMOS_3 had good correlation values for lower bitrate levels (i.e., higher levels of compression).

Inspired by these subjective audio-visual models, we combine a set of well-known audio and video quality metrics using all 3 combination models. This resulted in a set of FR and NR audio-visual quality metrics, which are described in the following sections.

5 FR audio-visual metrics

To design a FR audio-visual quality metric, we use 3 video quality metrics and 2 audio quality metrics. The chosen audio quality metrics are: the perceptual evaluation of audio quality (PEAQ) [23], a well-known standardized algorithm, and the virtual speech quality objective listener (VISQOL) [7], which has a good performance in comparis6on to other audio metrics [8, 9]. Additionally, both audio metrics are computationally inexpensive. Meanwhile, the chosen video metrics are: the video quality metric (VQM) [19], the peak signal-to-noise

[2] http://www.ene.unb.br/mylene/databases.html.

Table 3 Pearson correlation coefficients (PCC) of subjective models tested on low and high quality material sub-sets

Video bitrate (Mbps)	Audio bitrate (Kbps)	Number of Sequences	PCC $PrMOS_1$	PCC $PrMOS_2$	PCC $PrMOS_3$
Low (1, 0.8)	All (48, 96, 128)	36	0.8050	0.8178	0.8214
	Low (48, 96)	24	0.8227	0.8539	0.8540
	High (128)	12	0.6971	0.7268	0.7307
High (2, 30)	All (48, 96, 128)	36	0.8602	0.8769	0.8944
	Low (48, 96)	24	0.7891	0.8161	0.8441
	High (128)	12	0.9034	0.9119	0.8933
Global Results		**78**	**0.9110**	**0.9197**	**0.9285**

All values in bold type represent the best correlation results among the different values

ratio (PSNR), and the structural similarity (SSIM) index [26]. All three metrics are very well-known FR metrics, with relatively low computational complexity.

To obtain an audio-visual FR quality metric, the output of an audio metric and the output of a video metric are combined using one of the models described in Section 4. In total, 6 FR combination metrics (3 video × 2 audio) were tested for each model (linear, Minkowski, and power), resulting in 18 different combinations of FR metrics. Table 4 shows the Pearson and Spearman Correlation Coefficients (PCC and SCC, respectively) corresponding to the results of all 18 FR audio-visual combination metrics tested on the data of Experiment III. Additionally, correlation coefficients for the individual audio and video metrics are depicted at Table 5.

Notice that the audio and video metrics VISQOL and VQM have the best individual accuracy performances, reaching coefficient values around 0.40 and 0.70, respectively. The VQM-VISQOL combination metric has the best correlation coefficients, with values above 0.8 for all three models (linear, Minkowski, and power). In particular, the power model provides the best results (among all three models) with a PCC and SCC of 0.82 and 0.81, respectively. For the other combination metrics, a slightly better performance is obtained with the linear and power models. On the other hand, the PSNR-PEAQ and SSIM-PEAQ Minkowski combination metrics has the smallest correlation values. Analysing these

Table 4 Pearson and Spearman Correlation Coefficients (PCC and SCC) of the 18 FR audio-visual metrics – tested on UnB Audio-Visual Quality (UnB-AVQ) Database 1

Video	Audio	Linear		Minkowski		Power	
		PCC	SCC	PCC	SCC	PCC	SCC
VQM	VISQOL	**0.818**	**0.807**	**0.819**	**0.819**	**0.822**	**0.817**
	PEAQ	0.753	0.778	0.691	0.710	0.720	0.736
PSNR	VISQOL	0.750	0.741	0.745	0.730	0.757	0.749
	PEAQ	0.703	0.698	0.606	0.603	0.657	0.650
SSIM	VISQOL	0.707	0.704	0.667	0.664	0.710	0.720
	PEAQ	0.629	0.648	0.571	0.655	0.632	0.649

All values in bold type represent the best correlation results among the different values

Table 5 Pearson and Spearman Correlation Coefficients (PCC and SCC) of the individual FR audio and video metrics – tested on UnB Audio-Visual Quality (UnB-AVQ) Database 1

	Audio		Video		
Single Metric	VISQOL	PEAQ	VQM	PSNR	SSIM
PCC	0.424	−0.320	0.709	0.657	0.570
SCC	0.404	−0.321	0.736	0.651	0.662

results, we notice that a better integration capacity is achieved using the linear and power combination models.

To test if the differences in Table 4 are statistically significant, a two-tailed t-test was performed on the SCC values, considering 15 trials. These trials were set by randomly selecting 4 out of 6 original videos in the Database I and, then, calculating the SCC value. The SCCs values for each combination metric are then grouped and compared with each other. Figure 3 presents the box plot of the SCC values for each of the 18 FR audio-visual metrics.

T-test results for all FR combination metrics are presented in Table 6. Each cell in this table reports the null hypothesis test (95% confidence interval) between the pairs of mean correlation values of the combination metrics in the corresponding row and column. A cell value equal to “1” denotes that the performance of the row combination is statistically superior to the performance of the column combination, while a value “−1” denotes that the performance of the row combination metric is statistically worse than the performance of the column combination metric. Finally, a value of “0” denotes that both row and column combination metrics are statistically equivalent, in other words, the null hypothesis cannot be rejected.

From the results depicted in Table 6, the superior performance of the VQM-VISQOL combination metric, over all combination metrics, is confirmed. However, the results also show that there is no significant difference between the three models (linear, Minkowski, and power) for the VQM-VISQOL combination metric (t-Test results equal to “0”). Next,

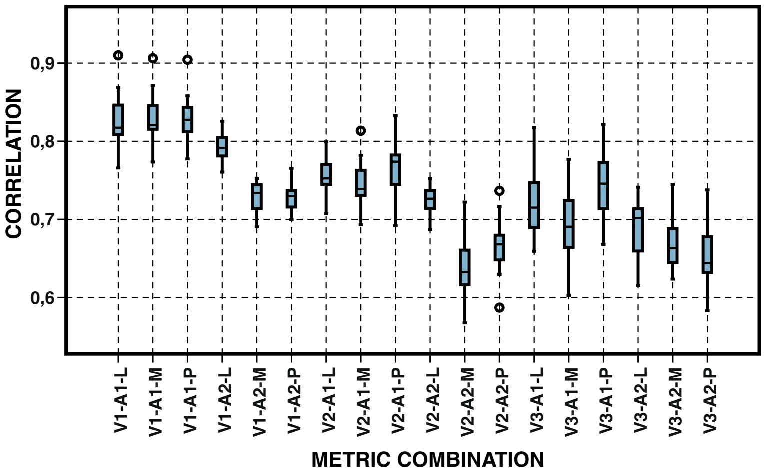

Fig. 3 Box plot of the SCC values of 18 audio-visual FR combination metrics, across 15 trials for UnB Audio-Visual Quality (UnB-AVQ) Database 1. Labels: V1 = VQM, V2 = PSNR, V3 = SSIM, A1 = VISQOL, A2 = PEAQ, L = Linear, M = Minkowski, P = Power

Table 6 Results of two-tailed t-Test executed on the SCC values obtained from 15 trials among the 18 FR audio-visual metrics for UnB Audio-Visual Quality (UnB-AVQ) Database 1

			VQM						PSNR						SSIM					
			VISQOL			PEAQ			VISQOL			PEAQ			VISQOL			PEAQ		
			L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P
VQM	VISQOL	L	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		M	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		P	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	PEAQ	L	−1	−1	−1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		M	−1	−1	−1	−1	0	0	−1	0	−1	0	1	1	0	1	0	1	1	1
		P	−1	−1	−1	−1	0	0	−1	0	−1	0	1	1	0	1	0	1	1	1
PSNR	VISQOL	L	−1	−1	−1	−1	1	1	0	0	0	1	1	1	1	1	0	1	1	1
		M	−1	−1	−1	−1	0	0	0	0	0	1	1	1	0	1	0	1	1	1
		P	−1	−1	−1	−1	1	1	0	0	0	1	1	1	1	1	0	1	1	1
	PEAQ	L	−1	−1	−1	−1	0	0	−1	−1	−1	0	1	1	0	1	0	1	1	1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	−1	0
		P	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	0	−1	0	0	0
SSIM	VISQOL	L	−1	−1	−1	−1	0	0	−1	1	−1	0	1	1	0	1	−1	1	1	1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	0	0	0	−1	0	0	1
		P	−1	−1	−1	−1	0	0	0	0	0	0	1	1	0	1	0	1	1	1
	PEAQ	L	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	0	−1	0	−1	0	0	1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	0	−1	0	−1	0	0	0
		P	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	0

Value “1” denotes row metric is superior to the column metric. Value “−1” denotes row metric worse to the column metric. Value of “0’ denotes both row and column metrics equivalent

the performance of the PSNR-VISQOL combination metric for the three models (linear, Minkowski, and power) and of the VQM-PEAQ combination metric for the linear model are superior to the performance of most of the other combination metrics. The weakest performance corresponded to the combination metrics PSNR-PEAQ and SSIM-PEAQ.

6 NR audio-visual metrics

The NR audio-visual metrics are obtained using 4 NR video quality metrics and 2 NR audio quality metrics. The chosen audio metrics are the original and reduced versions of the single ended speech quality assessment metric (SESQA and RSESQA) [12]. The SESQA metric, originally proposed for speech quality, and its reduced version RSESQA, both have a good accuracy performance for generic audio sequences [13]. Moreover, they are among the few NR Speech/Audio metrics currently available in the literature. Meanwhile, the chosen NR video metrics are: a blockiness-blurriness (BB) metric [25], the blind/referenceless image spatial quality evaluator (BRISQUE) [15], the blind image quality index (BIQI) [17], and the naturalness image quality evaluator (NIQE) [16]. These metrics were selected due to their low computational complexity and their good accuracy performance.

Table 7 Pearson and Spearman Correlation Coefficients (PCC and SCC) of NR audio-visual combination metrics – tested on UnB Audio-Visual Quality (UnB-AVQ) Database 1

Video	Audio	Linear		Minkowski		Power	
		PCC	SCC	PCC	SCC	PCC	SCC
BB	RSESQA	**0.793**	**0.797**	**0.778**	**0.792**	**0.810**	**0.807**
	SESQA	0.614	0.678	0.614	0.676	0.646	0.676
BRISQUE	RSESQA	0.541	0.494	0.544	0.504	0.537	0.453
	SESQA	0.379	0.324	0.365	0.283	0.407	0.317
BIQI	RSESQA	0.549	0.511	0.582	0.571	0.511	0.478
	SESQA	0.413	0.430	0.413	0.423	0.504	0.500
NIQE	RSESQA	0.541	0.494	0.543	0.506	0.540	0.456
	SESQA	0.379	0.324	0.369	0.291	0.408	0.333

All values in bold type represent the best correlation results among the different values

The outputs of an audio metric and a video metric (both NR) are combined using all combination models described on Section 4. A total of 8 NR combination metrics (4 video $\times$ 2 audio) were tested using the three combination models (linear, Minkowski, and power), what produced 24 different NR audio-visual quality combination metrics. Table 7 shows the PCCs and SCCs for all 24 NR audio-visual combination metrics tested on the data the audio-visual UnB Audio-Visual Quality (UnB-AVQ) Database 1. Results show that the BB-RSESQA combination metric presents the best performance. For this metric, the power model has a slightly better performance (PCC $=$ 0.81) when compared to the other two models. There is no clear performance superiority among the three models, but the power model has a slight advantage. Combinations metrics BRISQUE-SESQA and NIQE-SESQA presented the lowest correlation values.

The correlation coefficients corresponding to the individual performance of all audio and video metrics are shown at Table 8. These correlation values show that the proposed combination models are able to significantly improve the quality prediction. In fact, an analysis of all correlation coefficients indicates that all combination models improved the performance, with the power model presenting a slightly better integration capacity.

Again, two-tailed t-test was performed to determine whether the differences in correlation values between pairs of combination metrics are statistically significant. Here, we used the same parameters and methodology used for the set of FR combination metrics. Figure 4 shows the box plot of the SCC values for each of the 24 NR audio-visual combination metrics. T-test results for all NR combination metrics are presented in Table 9.

Results in Table 9 confirm that the BB-RSESQA combination metric has the best performance among all combination metrics. Yet, the differences in correlation among the

Table 8 Pearson and Spearman Correlation Coefficients (PCC and SCC) of the individual NR audio and video metrics – tested on UnB Audio-Visual Quality (UnB-AVQ) Database 1

	Audio		Video			
Single Metric	RSESQA	SESQA	BB	BRISQUE	BIQI	NIQE
PCC	0.432	0.132	0.614	0.317	0.306	0.317
SCC	0.380	0.280	0.670	0.290	0.328	0.289

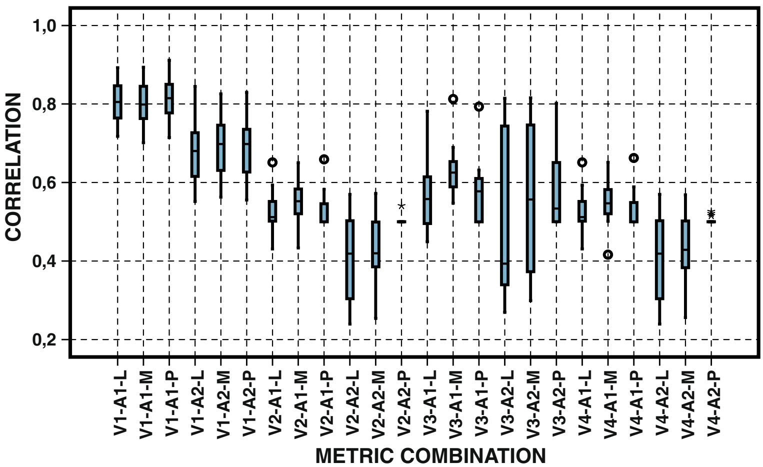

Fig. 4 Box plot of SCC values from 24 audio-visual NR metrics, across 15 trials in UnB Audio-Visual Quality (UnB-AVQ) Database 1. Labels: V1 = BB, V2 = BRISQUE, V3 = BIQI, V4 = NIQE, A1 = RSESQA, A2 = SESQA, L = Linear, M = Minkowski, P = Power

three combination models are not statistically significant. Surprisingly, using the Minkowski model for the BIQI-RSESQA combination metric results in a very good performance, only inferior to the performance of the BB-RSESQA and BB-SESQA combination metrics.

Table 9 Results of two-tailed t-Test executed on the SCC values obtained from 15 trials among the 24 NR audio-visual metrics in Database I

			BB						BRIS.						BIQI						NIQE					
			RSESQA			SESQA			RSESQA			SESQA			RSESQA			SESQA			RSESQA			SESQA		
			L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P
BB	RSESQA	L	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		M	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		P	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	SESQA	L	−1	−1	−1	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
		M	−1	−1	−1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		P	−1	−1	−1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRIS.	RSESQA	L	−1	−1	−1	−1	−1	−1	0	0	0	1	1	0	0	−1	0	0	0	0	0	0	0	1	1	0
		M	−1	−1	−1	−1	−1	−1	0	0	0	1	1	1	0	−1	0	0	0	0	0	0	0	1	1	1
		P	−1	−1	−1	−1	−1	−1	0	0	0	1	1	0	0	−1	0	0	0	0	0	0	0	1	1	0
	SESQA	L	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	0	−1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	0	−1
		P	−1	−1	−1	−1	−1	−1	0	−1	0	1	1	0	−1	−1	−1	0	0	−1	0	−1	−1	1	1	0
BIQI	RSESQA	L	−1	−1	−1	−1	−1	−1	0	0	0	1	1	1	0	−1	0	0	0	0	0	0	0	1	1	1
		M	−1	−1	−1	0	−1	−1	1	1	1	1	1	1	1	0	1	0	0	0	1	1	1	1	1	1
		P	−1	−1	−1	−1	−1	−1	0	0	0	1	1	1	0	−1	0	0	0	0	0	0	0	1	1	1
	SESQA	L	−1	−1	−1	−1	−1	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		M	−1	−1	−1	−1	−1	−1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0
		P	−1	−1	−1	−1	−1	−1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1
NIQE	RSESQA	L	−1	−1	−1	−1	−1	−1	0	0	0	1	1	0	0	−1	0	0	0	0	0	0	0	1	1	0
		M	−1	−1	−1	−1	−1	−1	0	0	0	1	1	1	0	−1	0	0	0	0	0	0	0	1	1	1
		P	−1	−1	−1	−1	−1	−1	0	0	0	1	1	1	0	−1	0	0	0	0	0	0	0	1	1	0
	SESQA	L	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	0	−1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	0	−1
		P	−1	−1	−1	−1	−1	−1	0	−1	0	1	1	0	−1	−1	−1	0	0	−1	0	−1	0	1	1	0

Value “1” denotes row metric is superior to the column metric. Value “−1” denotes row metric worse to the column metric. Value of “0’ denotes both row and column metrics equivalent

Fig. 5 Sample frames of the original videos of the NTIA Database [21]

Finally, the weakest performance corresponds to the BRISQUE-SESQA and NIQE-SESQA combination metrics.

7 NTIA audio-visual database analysis

Both sets of FR and NR audio-visual combination metrics were tested on a second database (**NTIA Database**), provided by The National Telecommunications and Information Administration (NTIA) [21]. This database contains sequences with audio and video components at VGA resolution (640 $\times$ 480, 4:2:2, 30 fps). For each original sequence, there are 5 test conditions, which correspond to different combinations of audio (8, 32, 64 KB/s) and video (100, 192, 250, 448, 500, 1000 KB/s) bitrates. Representative frames of the original videos are shown in Fig. 5. Each quality estimate obtained with all 18 FR and 24 NR audio-visual combination metrics are compared to the 10 subjective scores of the NTIA Database, which were gathered from 10 experiments performed in 6 different laboratories.

Table 10 Pearson and Spearman Correlation Coefficients (PCC and SCC) of the 18 FR audio-visual metrics – tested on NTIA Audio-Visual Database [21]

Video	Audio	Linear		Minkowski		Power	
		PCC	SCC	PCC	SCC	PCC	SCC
VQM	VISQOL	**0.520**	**0.544**	**0.521**	**0.543**	**0.522**	**0.530**
	PEAQ	0.402	0.412	0.407	0.425	0.412	0.468
PSNR	VISQOL	0.425	0.434	0.447	0.454	0.449	0.464
	PEAQ	0.265	0.283	0.298	0.293	0.301	0.372
SSIM	VISQOL	0.431	0.462	0.455	0.471	0.468	0.485
	PEAQ	0.219	0.248	0.251	0.259	0.302	0.501

All values in bold type represent the best correlation results among the different values

Table 11 Pearson and Spearman Correlation Coefficients (PCC and SCC) of the individual FR audio and video metrics – tested on NTIA Audio-Visual Database [21]

	Audio		Video		
Single Metric	VISQOL	PEAQ	VQM	PSNR	SSIM
PCC	0.285	0.132	0.241	0.203	0.242
SCC	0.351	0.250	0.253	0.209	0.245

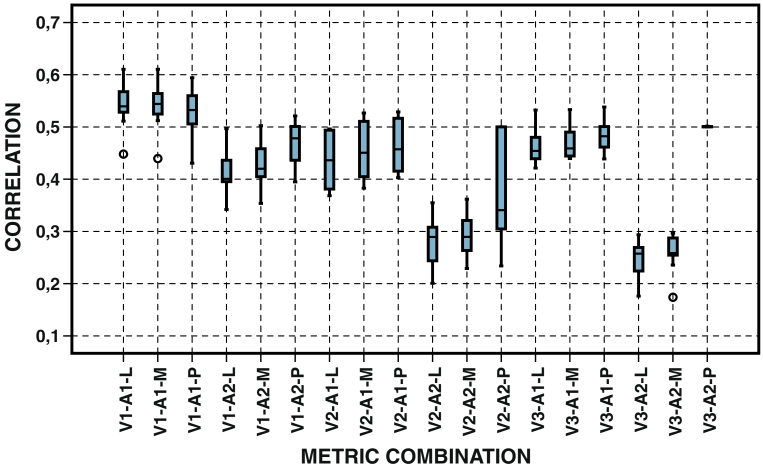

Fig. 6 Box plot of the SCC values for the 18 audio-visual FR combination metrics, tested on the NTIA audio-visual Database. Labels: V1 = VQM, V2 = PSNR, V3 = SSIM, A1 = VISQOL, A2 = PEAQ, L = Linear, M = Minkowski, P = Power

For the FR combination metrics, the average PCCs and SCCs obtained for all 10 datasets are shown in Table 10. Notice that all metrics have much lower correlation coefficients for this database, barely reaching 0.5. Although these results present lower correlations than the ones obtained for the UnB-AVQ Database 1, the VQM-VISQOL combination metric has a superior performance, in agreement with what was observed in our previous analysis. In fact, this combination has the best correlation values (between 0.52 and 0.54) for all three models (linear, Minkowski, and power), with no combination model standing out from the rest. Also, the PSNR-PEAQ and SSIM-PEAQ combinations presented the lowest correlations values, in agreement with the results observed for the UnB-AVQ Database 1.

The performance of the individual audio and video metrics are shown at Table 11. It is interesting to notice that VISQOL has a slightly better performance than the three other video quality metrics, although all the individual metrics have a very good performance. The analysis of the correlation values for the individual metrics and for their combination indicate that the three combination models provide a similar accuracy performance.

To verify whether the differences in correlation values are statistically relevant, a t-test was carried out. For this case, each of the 18 FR combination metrics produced a set of 10

Table 12 Results of two-tailed t-Test executed on the SCC values obtained from 10 subjective experiments (NTIA audio-visual Database) among the 18 FR audio-visual metrics

			VQM						PSNR						SSIM					
			VISQOL			PEAQ			VISQOL			PEAQ			VISQOL			PEAQ		
			L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P
VQM	VISQOL	L	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		M	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		P	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	PEAQ	L	−1	−1	−1	0	0	−1	0	0	−1	1	1	0	−1	−1	−1	1	1	−1
		M	−1	−1	−1	0	0	−1	0	0	0	1	1	0	0	−1	−1	1	1	−1
		P	−1	−1	−1	1	1	0	0	0	0	1	1	1	0	0	0	1	1	−1
PSNR	VISQOL	L	−1	−1	−1	0	0	0	0	0	0	1	1	0	0	0	−1	1	1	−1
		M	−1	−1	−1	0	0	0	0	0	0	1	1	0	0	0	0	1	1	−1
		P	−1	−1	−1	1	0	0	0	0	0	1	1	1	0	0	0	1	1	−1
	PEAQ	L	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	0	−1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	0	−1	−1	−1	1	0	−1
		P	−1	−1	−1	0	0	−1	0	0	−1	1	0	0	−1	−1	−1	1	1	−1
SSIM	VISQOL	L	−1	−1	−1	1	0	0	0	0	0	1	1	1	0	0	0	1	1	−1
		M	−1	−1	−1	1	1	0	0	0	0	1	1	1	0	0	0	1	1	−1
		P	−1	−1	−1	1	1	0	1	0	0	1	1	1	0	0	0	1	1	0
	PEAQ	L	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	0	−1
		M	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	−1	−1	−1	−1	0	0	−1
		P	−1	−1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0

Value "1" denotes that the row metric is superior to the column metric. Value "−1" denotes that the row metric worse to the column metric. Value of "0' denotes that both row and column metrics equivalent

Table 13 Pearson and Spearman Correlation Coefficients (PCC and SCC) of NR audio-visual combination metrics – tested on NTIA audio visual Database [21]

Video	Audio	Linear		Minkowski		Power	
		PCC	SCC	PCC	SCC	PCC	SCC
BB	RSESQA	0.735	**0.740**	**0.754**	**0.756**	**0.758**	**0.760**
	SESQA	**0.741**	0.714	0.741	0.704	0.743	0.704
BRISQUE	RSESQA	0.412	0.430	0.465	0.433	0.449	0.404
	SESQA	0.412	0.424	0.488	0.508	0.388	0.473
BIQI	RSESQA	0.476	0.479	0.527	0.460	0.468	0.454
	SESQA	0.464	0.459	0.484	0.508	0.451	0.460
NIQE	RSESQA	0.387	0.403	0.448	0.431	0.452	0.381
	SESQA	0.397	0.427	0.451	0.488	0.391	0.452

All values in bold type represent the best correlation results among the different values

correlation scores, which resulted from the comparison of the predicted quality and the subjective score gathered in each of the experiments. These correlation scores were grouped and used in a two-tailed t-test (95% confidence interval). Figure 6 shows the box plot of the SCC values of each of the 18 FR combination metrics, tested on Database II. Table 12 shows the t-test results of all these FR combination metrics. The VQM-VISQOL combination metric presents the best performance. Additionally, all three models of the combination metrics PSNR-VISQOL and SSIM-VISQOL exhibit a superior performance when compared to the other combination metrics. Moreover, the power model for the SSIM-PEAQ combination metrics also present a good performance. In summary, although these results are in agreement with the ones obtained for the UnB-AVQ Database 1 (see Section 5), but they show a considerable drop in the correlation values.

The set of NR audio-visual metrics was also tested on the NTIA audio-visual Database. Table 13 shows the average PCCs and SCCs for this database. A simple analysis suggests that the BB-RSESQA and BB-SESQA combination metrics performed much better than the rest of the combination metrics (PCC and SCC above 0.70). As for the combination models, a small advantage is observed for the Minkowski model. For this particular database, it is not possible to determine which audio metric has the better performance, but it is clear that the BB has the best performance among the video metrics. In a more global analysis, these results are (surprisingly) better than the results obtained for the FR metrics (see Table 10).

Analysing the individual performance of the metrics (Table 14), we observe that there is no considerable difference between the performance of the audio metrics RSESQA and

Table 14 Pearson and Spearman Correlation Coefficients (PCC and SCC) of the individual NR audio and video metrics – tested on NTIA audio visual Database [21]

	Audio		Video			
Single Metric	RSESQA	SESQA	BB	BRISQUE	BIQI	NIQE
PCC	0.364	0.357	0.633	0.052	0.172	0.010
SCC	0.360	0.390	0.619	−0.001	0.123	−0.063

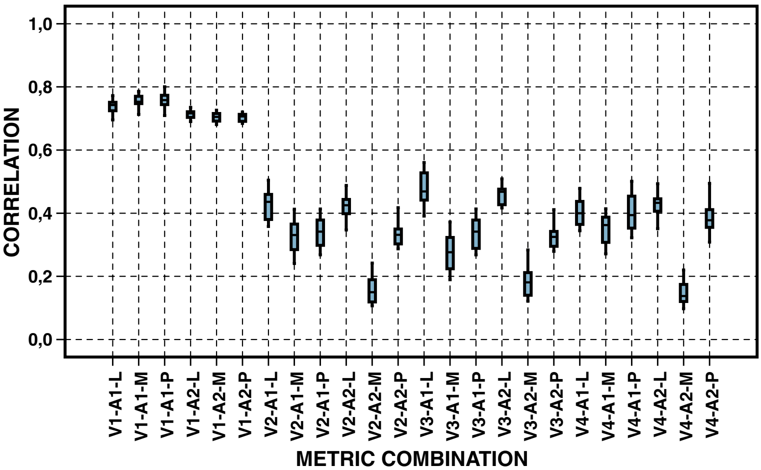

Fig. 7 Box plot of the SCC values for the 24 audio-visual NR combination metrics, tested on the NTIA Database. Labels: V1 = BB, V2 = BRISQUE, V3 = BIQI, V4 = NIQE, A1 = RSESQA, A2 = SESQA, L = Linear, M = Minkowski, P = Power

SESQA. Regarding the video metrics, there is a substantial gap between the performance of the BB metric and the performance of the rest of the video quality metrics. In terms of integration capacity, all three models presented a similar accuracy performance, with the Minkowski model performing slightly better.

Table 15 Results of two-tailed t-Test executed on the SCC values obtained from 10 subjective experiments (NTIA audio-visual Database) among the 24 NR audio-visual metrics

		BB						BRIS.						BIQI						NIQE					
		RSESQA			SESQA			RSESQA			SESQA			RSESQA			SESQA			RSESQA			SESQA		
		L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P	L	M	P
BB	RSESQA L	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	M	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	P	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	SESQA L	−1	−1	−1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	M	−1	−1	−1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	P	−1	−1	−1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRIS.	RSESQA L	−1	−1	−1	−1	−1	−1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0
	M	−1	−1	−1	−1	−1	−1	−1	0	0	−1	1	0	−1	0	0	−1	1	0	−1	0	−1	−1	1	−1
	P	−1	−1	−1	−1	−1	−1	−1	0	0	−1	1	0	−1	1	0	−1	1	0	−1	0	−1	−1	1	0
	SESQA L	−1	−1	−1	−1	−1	−1	0	1	1	0	1	1	−1	1	1	−1	1	1	0	1	0	0	1	0
	M	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	−1
	P	−1	−1	−1	−1	−1	−1	−1	0	0	−1	1	0	−1	1	0	−1	1	0	−1	0	−1	−1	1	−1
BIQI	RSESQA L	−1	−1	−1	−1	−1	−1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1
	M	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	1	−1	−1	0	−1	−1	1	−1	−1	−1	−1	−1	1	−1
	P	−1	−1	−1	−1	−1	−1	−1	0	0	−1	1	0	−1	1	0	−1	1	0	−1	0	−1	−1	1	0
	SESQA L	−1	−1	−1	−1	−1	−1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
	M	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	−1
	P	−1	−1	−1	−1	−1	−1	−1	0	0	−1	1	0	−1	1	0	−1	1	0	−1	0	−1	−1	1	−1
NIQE	RSESQA L	−1	−1	−1	−1	−1	−1	0	1	1	0	1	1	−1	1	1	−1	1	1	0	1	0	0	1	0
	M	−1	−1	−1	−1	−1	−1	−1	0	0	−1	1	0	−1	1	0	−1	1	0	−1	0	−1	−1	1	0
	P	−1	−1	−1	−1	−1	−1	0	1	1	0	1	1	−1	1	1	−1	1	1	0	1	0	0	1	0
	SESQA L	−1	−1	−1	−1	−1	−1	0	1	1	0	1	1	−1	1	1	0	1	1	0	1	0	0	1	0
	M	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	−1	−1	−1	−1	−1	0	−1
	P	−1	−1	−1	−1	−1	−1	0	1	0	0	1	1	−1	1	0	−1	1	1	0	0	0	0	1	0

Value “1” denotes row metric is superior to the column metric. Value “−1” denotes row metric worse to the column metric. Value of “0’ denotes both row and column metrics equivalent

A two-tailed t-test was carried out in order to verify the significance of the differences among the correlation values obtained by the NR audio-visual metrics for the 10 subjective experiments from NTIA audio-visual Database. A box plot of the SCCs scores for all 24 NR audio-visual metrics is depicted in Fig. 7. Table 15 presents the results of this t-test. Notice that the BB-RSESQA and BB-SESQA combination metrics have a superior performance for all three models (linear, Minkowski, and power). Moreover, among the remaining NR metrics, for most of combination metrics, a better performance is obtained for the linear model.

8 Conclusions

In this work, we studied the use of combination models to integrate single audio and video quality estimates with the goal of predicting the overall audio-visual quality. To obtain the audio and video quality estimates, we used a set of mature and sufficiently tested audio and video quality metrics, considering both FR and NR approaches. For the FR approach, we chose 3 video quality metrics (VQM, PSNR, and SSIM) and 2 audio quality metrics (VISQOL and PEAQ), while for the NR approach, we chose 4 video quality metrics (BB, BRISQUE, BIQI, NIQE) and 2 audio quality metrics (RSESQA and SESQA). The individual predictions of audio and video quality metrics were integrated using three combination models: linear, Minkowski, and power. The audio and video metrics were combined and this resulted in 18 FR and 24 NR audio-visual quality metrics.

All 18 FR and 24 NR metrics were tested on two different audio-visual databases. For the FR type of metric, a considerable difference of the correlations is observed between the two databases under study (above 0.8 for UnB-AVQ Database I and 0.5 for NTIA audio-visual Database). Nevertheless, the VQM-VISQOL combination metric presented the best results for both databases. This combination metric performed well for all three combination functions, with a small advantage of the power model. It was also observed that metrics like PSNR, SSIM (video) and PEAQ (audio) did not perform very well on any of the databases. Meanwhile, for the NR audio-visual metrics, the BB-RSESQA and BB-SESQA combinations presented a superior performance for both databases. These combination metrics presented an equivalent performance for the three models (linear, Minkowski, and power). On the other hand, video quality metrics like BRISQUE and NIQE did not perform well in any of the databases.

Observing the performance of the individual metrics, we noticed that the three combining models have a good integration capacity. It is worth pointing out that, out of the three models, only the linear model was previously used for combining audio and video objective scores. Therefore, one of the goals of this work was to test different combination models and study their integration capacity, in terms of accuracy performance. Although the results are promising, we believe an improvement in performance can be obtained by taking into account the interaction between the human visual and auditory systems. Also, better performance can be achieved using more complex combination models (e.g. machine learning based algorithms). Finally, we need to perform tests using more diverse audio-visual databases, containing several types of audio and video degradations. Unfortunately, up to our knowledge, this type of database is not currently available.

Acknowledgments This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Brazil, in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brazil, and in part by the University of Brasília.

References

1. Bong DBL, Khoo BE (2015) Objective blur assessment based on contraction errors of local contrast maps. Multimed Tools Appl 74.17:7355–7378
2. Chikkerur S, Sundaram V, Reisslein M, Karam LJ (2011) Objective video quality assessment methods: a classification, review, and performance comparison. IEEE Trans Broadcast 57(2):165–182
3. Engelke U, Zepernick H-J (2007) Perceptual-based quality metrics for image and video services: a survey. In: 3rd EuroNGI Conference on next generation internet networks, pp 190–197
4. Farias MCQ, Mitra SK (2005) No-reference video quality metric based on artifact measurements. IEEE Int Conf Image Process ICIP 2005 3(2):III–141–4
5. Garcia MN, Schleicher R, Raake A (2011) Impairment-factor-based audiovisual quality model for iptv Influence of video resolution, degradation type, and content type. EURASIP J Image Vid Process 1–14
6. Hands DS (2004) A basic multimedia quality model. IEEE Trans Multimed 6(6):806–816
7. Hines A, Skoglund J, Kokaram A, Harte N (2012) Visqol: the virtual speech quality objective listener. In: International workshop on acoustic signal enhancement. Proceedings of IWAENC 2012. VDE, pp 1–4
8. Hines A, Skoglund J, Kokaram A, Harte N (2013) Robustness of speech quality metrics to background noise and network degradations: comparing visqol, pesq and polqa. In: 2013 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3697–3701
9. Hines A, Gillen E, Harte N (2015) Measuring and monitoring speech quality for voice over IP with POLQA, ViSQOL and P. 563. In: Interspeech conference. Dresden, pp 6–10
10. ITU Recommendation BT.500-8 (1998) Methodology for subjective assessment of the quality of television pictures
11. Lin W, Kuo C-CJ (2011) Perceptual visual quality metrics: a survey. J Vis Commun Image Represent 22(4):297–312
12. Malfait L, Berger J, Kastner M (2006) P.563 The itu-t standard for single-ended speech quality assessment. IEEE Trans Audio Speech Lang Process 14(6):1924–1934
13. Martinez HB, Farias MCQ (2014) A no-reference audio-visual video quality metric. In 2014 Proceedings of the 22nd European on signal processing conference (EUSIPCO), pp 2125–2129
14. Martinez HB, Farias MCQ (2014) Full-reference audio-visual video quality metric. J Electron Imag 23(6):061108–061108
15. Mittal A, Moorthy AK, Bovik AC (2011) Blind/referenceless image spatial quality evaluator. In: Conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR). IEEE, pp 723–727
16. Mittal A, Soundararajan R, Bovik AC (2013) Making a “completely blind” image quality analyzer. IEEE Signal Process Lett 20(3):209–212
17. Moorthy AK, Bovik AC (2010) A two-step framework for constructing blind image quality indices. IEEE Signal Process Lett 17(5):513–516
18. Moorthy AK, Bovik AC (2011) Visual quality assessment algorithms: what does the future hold? Multimed Tools Appl 51(2):675–696
19. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322
20. Pinson MH, Ingram W, Webster A (2011) Audiovisual quality components. Signal Process Mag IEEE 28(6):60–67
21. Pinson MH, Schmidmer C, Janowski L, Pépion R, Huynh-Thu Q, Corriveau P, Younkin A, Le Callet P, Barkowsky M, Ingram W (2013) Subjective objective evaluation of an audiovisual subjective dataset for research and development. In: Fifth International workshop on quality of multimedia experience (QoMEX). IEEE, pp 30–31
22. Soh K, Iah S (2001) Subjectively assessing method for audiovisual quality using equivalent signal-to-noise ratio conversion. Trans Inst Electron Inform Commun Eng A 11:1305–1313
23. Thiede T, Treurniet WC, Bitto R, Schmidmer C, Sporer T, Beerends JG, Colomes C (2000) Peaq-the itu standard for objective measurement of perceived audio quality. J Audio Eng Soc 48(1/2):3–29
24. Wang Z, Sheikh HR, Bovik A (2002) No-reference perceptual quality assessment of jpeg compressed images. IEEE Int Conf Proc 1:I–477–I–480
25. Wang Z, Sheikh HR, Bovik AC (2002) No-reference perceptual quality assessment of jpeg compressed images. In: 2002 International conference on image processing. 2002 Proceedings, vol 1, pp I–477–I–480
26. Wang Z, Lu L, Bovik A (2004) Video quality assessment based on structural distortion measurement. Signal Process Image Comm 19:121–132
27. Winkler S, Faller C (2006) Perceived audiovisual quality of low-bitrate multimedia content. IEEE Trans Multimed 8(5):973–980
28. Yamagishi K, Gao S (2013) Light-weight audiovisual quality assessment of mobile video Itu-t rec. p.1201.1. In: IEEE 15th International workshop on multimedia signal processing (MMSP), pp 464–469

Helard A. Becerra Martinez is a Ph.D. student at the Department of Computer Science of the University of Brasília. He received his BS degree in Computer Science in 2010 from the Universidad Nacional San Antonio Abad del Cusco (UNSAAC), Peru, and his M.Sc. degree in Computer Science in 2013 from the University of Brasília, Brazil. Mr. Becerra 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 B.Sc. degree in electrical engineering from Universidade Federal de Pernambuco (UFPE), Brazil, in 1995 and her M.Sc. degree in electrical engineering from the Universidade Estadual de Campinas (UNICAMP), Brazil, in 1998. She received her Ph.D. in electrical and computer engineering from the University of California Santa Barbara, USA, in 2004 for work in no-reference video quality metrics. Dr. Farias has worked as a research engineer at CPqD (Brazil) in video quality assessment and validation of video quality metrics. She has also worked for Philips Research Laboratories (The Netherlands) in video quality assessment of sharpness algorithms and for Intel Corporation (Phoenix, USA) developing no-reference video quality metrics. Currently, she is an Assistant Professor in the Department of Electrical Engineering at the University of Brasília (UnB), where she is a member of the Graduate Program in Informatics and of the Graduate Program on Electronic Systems and Automation Engineering (PGEA). Dr. Farias is a researcher of the Group of Digital Signal Processing (GPDS) of UnB and her current interests include video quality metrics, video processing, multimedia signal processing, watermarking, and information theory. Dr. Farias is a member of IEEE and of the IEEE Signal Processing Society.