A NO-REFERENCE AUDIO-VISUAL VIDEO QUALITY METRIC

Helard Becerra Martinez[†] *and Mylène C. Q. Farias*[*†]

Department of Electrical Engineering[*]
Department of Computer Science[†]
University of Brasília, Brasília - DF, Brazil

ABSTRACT

Three psychophysical experiments were carried out to understand both audio and video components interact and affect the overall audio-visual quality. In the experiments, subjects independently evaluated the perceived quality of (1) video (without audio), (2) audio (without video), and (3) video with audio. With the help of the perceptual models obtained using subjective data, we propose 3 no-reference audio-visual quality metrics composed of combination functions of a video and an audio quality metrics. The no-reference video quality metric consists of a blockiness and a blurriness metrics, while the NR audio metric is modification of the SESQA metric. When tested on our database and on a public database, the metrics performed better than single video NR and RF metrics available in the literature.

***Index Terms*—** video quality assessment, quality metrics, audio-visual.

1. INTRODUCTION

Multimedia communication has evolved into an important field in the past few years. There have been significant advances in compression and transmission techniques, which have made possible to deliver high quality content to the end user. In particular, the advent of new technologies has allowed the creation of services like direct broadcast satellite, digital television, high definition TV, and Internet video. The level of acceptability and popularity of these services is related to its reliability and to the quality of the content provided. As a consequence, the development of real-time quality monitoring schemes is key for the success of these and future services.

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 full-reference (FR) video quality metrics [1]. Much remains to be done in the area of no-reference (NR) quality metrics [2]. Also, very few objective metrics have addressed the issue of simultaneously measuring the quality of all media involved (e.g. video, audio, text). 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 perceptual models [3, 4]. A good review on audio-visual quality studies was performed by Pinson *et al.* [5].

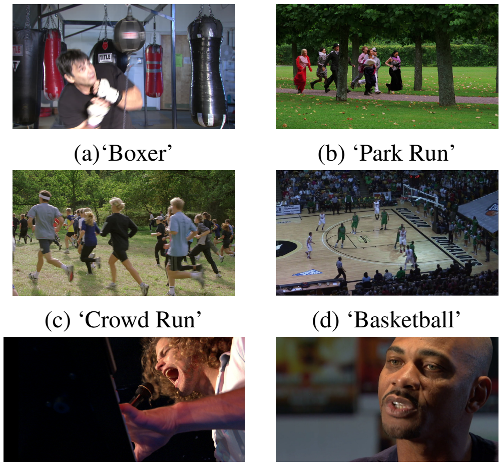
There are fewer works that tackle the development of audio-visual objective metrics. Among the most relevant works, we can cite the parametric (NR) objective quality metrics proposed by Garcia *et al.* [4] and Yamagishi and Gao [6]. Parametric metrics estimate quality using the information available at the receiver, such as bitrate, frame rate, QP, motion vectors, and various types of information regarding the impacts of network impairments. These metrics are generally faster than pixel-based video quality metrics and, depending on the level of access to the bitstream, can produce reliable results. However, parametric metrics are coding and transmission dependent, what makes them less generally applicable. In other words, they cannot predict the quality of content outside the 'transmission scenario', like, for example, content transcoded among different compression standards/bitrates or processed using any signal processing techniques. In this work, we are interested in developing generic pixel-based audio-visual quality metrics that can be used in most applications.

Previously [7], we have proposed a FR audio-visual quality metric based on a combination of FR audio and video quality metrics. In this work, our goal is to design a NR objective metric for audio-visual quality using a similar approach. To achieve this goal, we use the same experimental data from our previous work, which consisted of three psycho-physical experiments that independently measured the perceived quality of (1) video (without audio), (2) audio (without video), and (3) video with audio. In this work, we perform a new analysis on the experimental data and propose a set of perceptual audio-visual quality models. The perceptual models are the basis of the proposed NR objective metric, which consists of an audio metric, a blockiness metric, and a blurriness metric.

2. PSYCHOPHYSICAL EXPERIMENTS

For all three experiments, we used a set of six videos of eight seconds, obtained from The Consumer Digital Video

This work was supported in part by the University of Brasília and in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).



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

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

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

Fig. 1. Sample frames of videos used in experiments.

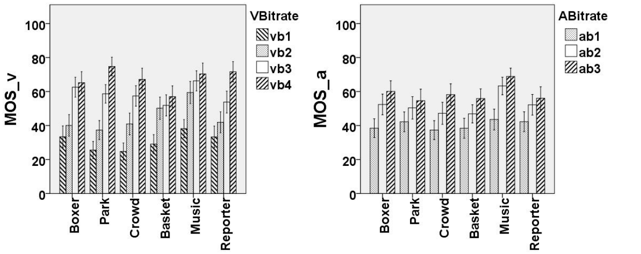
Table 1. Test Conditions Used in the Experiments I-III.

	Experiment I	Experiment II	Experiment III
Component	Video	Audio	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

Library (CDVL) website (http://www.cdvl.org/). Representative frames are shown in Figure 1. The videos are 1280x720 (4:2:0) and 30 frames per second (fps). We selected sequences that had speech, music, and ambient sound [7]. The ffmpeg framework was used to encode the test sequences, using different levels of bitrate levels for audio and video components. Details of the codecs, bitrates, number of sequences, and number of experimental subjects are listed at Table 1.

A Double-Stimulus Continuous Quality-Scale (DSCQS) methodology was used in all experiments [8], which consists of presenting two sequences (reference and test) with the same content and asking participants to score the quality of both of them. The scale used ranged from '0' to '100' and presentation order is random. For each test sequence, the Mean Opinion Score (MOS) is calculated by taking the average of the scores over all participants. Different groups of subjects were used in each experiment.

In Experiment I, 16 participants scored video test sequences (without audio), generating one MOS_v value for each test sequence. Figure 2(a) shows the MOS_v versus the video bitrate (vb) values (vb1 = 800 Kbps, vb2 = 1 Mbps, vb3 = 2 Mbps, vb4 = 30 Mbps) for all test sequences. One of the lowest MOS_v values (at high bitrate) correspond to the sequence 'Basketball' (low temporal and spatial activity). In contrast, two of the highest MOS_v values (at high bitrate) correspond to the videos 'Music' and 'Crowd Run' (high temporal and spatial activities). Due to the scene's characteristics and masking properties, some impairments might not be perceived by users. In other words, errors of the same type and the same energy level (mean squared error) when presented in complex scenes have a higher visibility threshold than when present in lower activity scenes.



(a) Experiment I (b) Experiment II

Fig. 2. (a) MOS_v (Exp.I) versus video bitrate and (b) MOS_a (Exp.II) versus audio bitrate.

In Experiment II, 16 subjects scored audio test sequences (without video), generating one MOS_a value for each test sequence. Figure 2.(b) shows the MOS_a versus the audio bitrate values (ab1 = 48 kbps, ab2 = 96 kbps, ab3 = 128 kbps) for all test sequences. The 'Basketball' audio sequence (environmental sounds) presented a slightly lower MOS value (not statistically significant). Meanwhile, the audio sequences 'Music' and 'Park Run' (Music, Screams, and Others2) got slightly higher MOS_a values.

In Experiment III, three audio bitrates and four video bitrates were used (same bitrates of Experiments I and II). 17 subjects performed the experiment, providing one MOS_{av} for each audio-visual test sequence. Figure 3 shows MOS_{av} versus the audio and video bitrates. For comparison, in the graphs we also show the data of Experiment I (video with no audio), which is represented as 'ab0'.

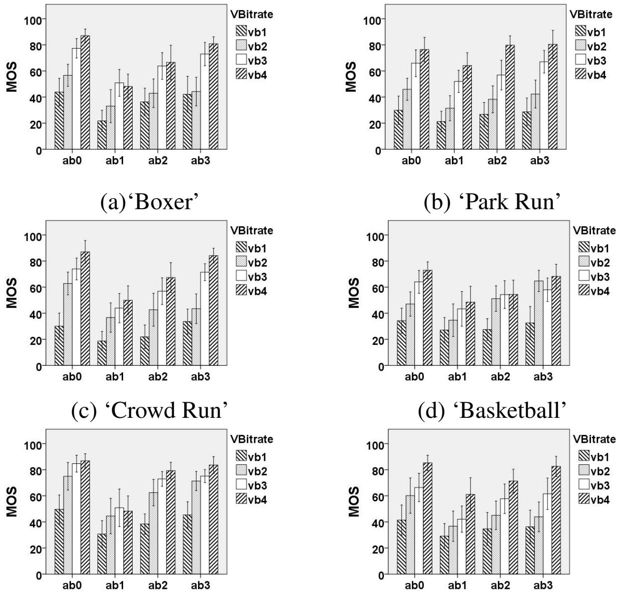
It can be observed from Figure 3 that the MOS_{av} values increase as the video bitrate values increase. Nevertheless, the slope caused by the increase in video bitrate is *not* the same for the different 'originals' or the different groups of audio bitrates (ab). This can be observed for the sequences 'Boxer', 'Basketball' and 'Music', which have different slopes for the different audio bitrates. Meanwhile, the sequences 'Park Run', 'Crowd Run', and 'Reporter' maintain similar slopes. In general, by comparing the results in all three experiments, we can observe that the video component had considerable more influence over the overall audio-visual quality.

3. PERCEPTUAL 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 main intention is to see which model integrates better the audio and video quality values.

The first perceptual model was a simple linear model:

$$\text{PrMOS}_1 = \alpha_1 \cdot \text{MOS}_v + \beta_1 \cdot \text{MOS}_a + \gamma_1. \quad (1)$$



(a)‘Boxer’ (b) ‘Park Run’

(c) ‘Crowd Run’ (d) ‘Basketball’

(e) ‘Music’ (f) ‘Reporter’

Fig. 3. Exp. I and III: MOS$_v$ and MOS$_{av}$ versus audio (and video) bitrates.

The fitting returned scaling coefficients $\alpha_1 = 0.76$, $\beta_1 = 0.41$, and $\gamma_1 = -21.92$.

The second model was a weighted Minkowski function:

$$\mathrm{PrMOS}_2 = (\alpha_2 \cdot \mathrm{MOS}_v^{p_1} + \beta_2 \cdot \mathrm{MOS}_a^{p_1})^{\frac{1}{p_1}}. \qquad (2)$$

The fit returned $p_1 = 0.0001$, $\alpha_2 = 0.7024$, and $\beta_2 = 0.2976$.

The last perceptual model tested was a power model:

$$\mathrm{PrMOS}_3 = (\gamma_2 + \alpha_3 \cdot \mathrm{MOS}_v^{p_2} \cdot \mathrm{MOS}_a^{p_3}), \qquad (3)$$

The fit returned $p_2 = 1.3213$, $p_3 = 0.6533$, $\alpha_3 = -0.0109$, and $\gamma_2 = -12.9734$.

We compared the perceptual models obtained in this section with three perceptual models available in the literature: two models (SQav$_{\mathrm{H1}}$ and SQav$_{\mathrm{H2}}$) proposed by Hands [3], two models (SQav$_{\mathrm{W1}}$ and SQav$_{\mathrm{W1}}$) proposed by Winkler [9], and one model (SQav$_{\mathrm{G}}$) proposed by Garcia [4]. Table 2 depicts the Pearson and Spearman correlation coefficients obtained by testing all perceptual models in the data of Experiment III. As can be observed, for this database, the proposed models present better results. But, models taken from literature present an acceptable correlation, given that they were not trained on this dataset.

4. OBJECTIVE QUALITY METRIC (NR)

To generate the audio-visual quality NR quality metric, we combined an audio and a video NR metrics. For the audio metric, we chose the NR speech quality metric SESQA (Single Ended Speech Quality Assessment Model) [10]. For the video metric, we combined two NR artifact metrics: a blurriness metric [11] and a blockiness metric [12]. These metrics were fast metrics that showed a good performance in tests using additional video and audio databases.

Table 2. Perceptual Audio-Visual Models: Pearson and Spearman Correlation Coefficients obtained for data of Exper. III.

Model	Type	Pearson	Spearman
PrMOS$_1$	SUBJ.	0.9110	0.9173
PrMOS$_2$	SUBJ.	0.9197	0.9267
PrMOS$_3$	SUBJ.	**0.9285**	**0.9270**
SQav$_{\mathrm{H1}}$	SUBJ.	0.8447	0.8340
SQav$_{\mathrm{H2}}$	SUBJ.	0.8441	0.8349
SQav$_{\mathrm{G}}$	SUBJ.	0.7739	0.8050
SQav$_{\mathrm{W1}}$	SUBJ.	0.8441	0.8349
SQav$_{\mathrm{W2}}$	SUBJ.	0.8244	0.8374

4.1. Audio Quality Metric

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

In order to apply this metric for audio signals (speech, music, generic sounds, etc.), we modified it slightly. Instead of using the 51 parameters considered in the original algorithm, we selected 17 parameters that showed better results 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 3. The rest of the SESQA algorithm was kept without modifications.

Table 3. Selected 17 SESQA parameters for audio metric [10].

Parameter	Name	Classification
1-2	PitchAverage, SpeechLevel	Basic voice descriptors
3	MuteLength	Interruptins/mutes
4-9	LocalBGNoiseLog, RelNoiseFloor, SNR, SpecLevelDev SpecLevelRange, SpectralClarity	Noise analysis
10-17	BasicVoiceQuality, ArtAverage CepCurt,FinalVtpAverage, LPCCurt LPCSkew, PitchCrossCorrelOffset PitchCrossPower	Unnatural speech

4.2. Video Quality Metric

The proposed NR video quality metric is composed by two artifact metrics: a blurriness metric proposed by Narvekar and Karem [11] and a blockiness metric proposed by Wang and Bovik [12]. For the blockiness metric, the algorithm calculates the vertical and horizontal absolute differences of the intensities fo the video frame. Then, blockiness is calculated by observing the peaks at the frequencies 1/8, 2/8, 3/8, and

4/8. Vertical and horizontal values are combined to obtain an estimate of blurriness for the video.

The blurriness metric uses the concept of just-noticeable blurriness together with a cumulative probability of blurriness detection. In other words, it uses the sensitivity of human blurriness perception at different contrast levels to estimate the probability of blurring being detected at each (strong) edge of the video frame. By evaluating the cumulative probability of blurriness detection, the blurriness perception information of each edge is calculated and summed over the entire frame to get a final blockiness score.

The NR video quality metric is obtained by combining the blurriness and blockiness scores using a simple linear model, given by the following equation:

$$\mathrm{Qv} = -195.08 \cdot \mathrm{Blur} + -55.23 \cdot \mathrm{Block} + 320.94. \quad (4)$$

This metric was trained using different videos from the ones used in the experiments described in Section 2.

4.3. Proposed Audio-Visual Quality Metric

We propose three NR audio-visual quality metrics, which are based on the perceptual models described in Section 3. The first audio-visual metric is a simple linear model, given by the following equation:

$$\mathrm{Qav}_1 = \alpha_4 \cdot \mathrm{Qv} + \beta_3 \cdot \mathrm{Qa} + \gamma_3, \quad (5)$$

where Qav_1 corresponds to the resulting predicted audio-visual quality score, Qv to the quality score obtained with the video quality metric, and Qa to the quality score obtained with the audio metric. The fit returned $\alpha_4 = 0.87$, $\beta_3 = 0.52$, and $\gamma_3 = -35.6387$. For this fit, the Pearson correlation coefficient was 0.7929 and the Spearman correlation coefficient was 0.7972.

The second audio-visual metric uses a weighted Minkowski model, given by the following equation:

$$\mathrm{Qav}_2 = (\alpha_5 \cdot \mathrm{Qv}^{p_4} + \beta_4 \cdot \mathrm{Qa}^{p_4})^{\frac{1}{p_4}}. \quad (6)$$

where Qav_2 corresponds to the predicted audio-visual quality score. The fit for the Minkowski model returned an exponent $p_4 = 0.003$ and scaling coefficients $\alpha_5 = 0.6160$ and $\beta_4 = 0.3840$. For this fit, the Pearson correlation coefficient was 0.7779 and the Spearman correlation coefficient was 0.7920.

Finally, the third audio-visual metric is a power model, given the following equation:

$$\mathrm{Qav}_3 = (\gamma_4 + \alpha_6 \cdot \mathrm{Qv}^{p_5} \cdot \mathrm{Qa}^{p_6}), \quad (7)$$

where Qav_3 corresponds to the predicted audio-visual quality score. The fit for this model returned exponents $p_5 = 1.9904$ and $p_6 = 0.9762$ and scaling coefficients $\alpha_6 = 0.0001$ and $\gamma_4 = 20.1468$. For this fit, the Pearson correlation coefficient was 0.8100 and the Spearman correlation coefficient was 0.8068.

Table 4. Pearson and Spearman Correlation Coefficients for data on Experiment III.

Model	Type	Pearson	Spearman
Qav_1	NR	0.7929	0.7972
Qav_2	NR	0.7779	0.7920
Qav_3	NR	**0.8100**	**0.8068**
SSIM	FR	0.5896	0.6435
VQM	FR	0.7092	0.7364
PSNR	FR	0.5437	0.6350
NIQE	NR	0.3901	0.3976
BIQI	NR	0.3607	0.3021
BRISQUE	NR	0.5804	0.5610

Due to the difficulty of finding NR (pixel-based) audio-visual quality metrics, we compared the proposed metrics with a group of FR and NR *video* metrics. The FR video quality metrics considered here are: SSIM [13], VQM [14], and PSNR. The NR video metrics considered are: NIQE [15], BIQI [16], and BRISQUE [17]. In Table 4, we summarize the Pearson and Spearman correlation coefficients obtained testing these metrics and the proposed audio-visual metrics on the data of Experiment III. As can be observed, similarly to the perceptual models, the proposed audio-visual metrics have the best performance, with the power model achieving the best correlation with subjective scores.

We also tested this set of metrics using a database provided by The National Telecommunications and Information Administration (NTIA) [5, 18], which is available for download at CDVL (www.cdvl.org). The database consists of data gathered at six different international laboratories associated to VQEG, resulting in ten sets of audio-visual MOS values [18]. The database sequences contained audio and video, with VGA resolution (640x480, 4:2:2, 30 fps). The Pearson correlation coefficients for this test is shown in Table 5 for the 10 sets of test sequences. Notice that all models have much lower correlation coefficients for this database, but the proposed audio-visual metrics have better correlation values. This result is expected since the other metrics only take into account the video component. But, surprisingly, the proposed Minkowski combination presented the best results.

5. CONCLUSIONS

We carried out three psychophysical experiments with the goal of understanding how the audio and video components contribute to the overall audio-visual perceptual quality. Based on the collected experimental data, we obtained three perceptual audio-visual models: a linear model, a weighted Minkowski model, and a power model. We also tested other five perceptual audio-visual quality models proposed in literature. For the given database, the proposed power model presented the best results.

Using video and audio NR metrics, we were able to obtain three objective NR audio-visual quality metrics. The combination functions used by the NR audio-visual metrics were the same used for the perceptual models. When compared to

Table 5. Pearson Correlation Coefficients for NTIA database [18].

Exp.	SSIM	PSNR	VQM	NIQE	BIQI	BRISQUE	Linear	Power	Mink.
NTIA_lab	0.2622	0.3221	0.2555	0.5964	0.3726	0.1985	0.5392	0.6783	0.7378
NTIA caf.	0.2589	0.4183	0.2475	0.5600	0.3154	0.0830	0.5131	0.6501	0.7307
Intel	0.3084	0.3614	0.3057	0.6056	0.3739	0.2068	0.5814	0.6875	0.7390
IRCCyN BT500	0.2858	0.3469	0.2990	0.5913	0.3477	0.1964	0.6014	0.6977	0.7653
IRCCyN Tablet	0.2728	0.3149	0.3464	0.5597	0.3724	0.2717	0.6264	0.6954	0.7573
Technicolor Dark Room	0.3038	0.3865	0.3458	0.6165	0.4062	0.2478	0.5946	0.7082	0.7314
Technicolor Patio	0.3069	0.3763	0.3259	0.6050	0.3792	0.2227	0.5777	0.6722	0.7191
AGH Lab	0.3600	0.4101	0.2747	0.5843	0.4022	0.1885	0.5500	0.6657	0.7028
AGH D5	0.2926	0.3352	0.3808	0.6201	0.4059	0.2738	0.6084	0.6923	0.7335
Opticom	0.319	0.3199	0.3683	0.6312	0.4601	0.2870	0.5997	0.6892	0.7320

NR and FR *video* quality metrics, the proposed metrics presented a better performance on our database. When tested on a public database provided by NTIA, all metrics presented lower correlation values, but the proposed metric presented significant better results.

REFERENCES

[1] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, “Objective video quality assessment methods: A classification, review, and performance comparison,” *IEEE Journal BC*, vol. 57, no. 2, pp. 165–182, 2011.

[2] S. S. Hemami and A. R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469 – 481, 2010, Special Issue on Image and Video Quality Assessment.

[3] David S Hands, “A Basic Multimedia Quality Model,” *Multimedia, IEEE Trans. on*, vol. 6, no. 6, pp. 806–816, 2004.

[4] 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 Journal on Image and Video Processing*, pp. 1–14, 2011.

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

[6] K. Yamagishi and S. Gao, “Light-weight audiovisual quality assessment of mobile video: Itu-t rec. p.1201.1,” in *Multimedia Signal Processing (MMSP), IEEE 15th International Workshop on*, Sept 2013, pp. 464–469.

[7] H. B. Martinez and M. C. Q. Farias, “An objective model for audio-visual quality,” in *Proc. SPIE, Image Quality and System Performance XI*, 2014, vol. 9016, pp. 90160P–90160P–14.

[8] ITU Recommendation BT.500-8, “Methodology for subjective assessment of the quality of television pictures,” Tech. Rep., ITU, 1998.

[9] S. Winkler and C. Faller, “Perceived audiovisual quality of low-bitrate multimedia content,” *Multimedia, IEEE Trans. on*, vol. 8, no. 5, pp. 973–980, 2006.

[10] L. Malfait, J. Berger, and M. Kastner, “P.563 amp;8212;the itu-t standard for single-ended speech quality assessment,” *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 14, no. 6, pp. 1924–1934, 2006.

[11] N.D. Narvekar and L.J. Karam, “A no-reference image blur metric based on the cumulative probability of blur detection (cpbd),” *Image Processing, IEEE Trans. on*, vol. 20, no. 9, pp. 2678–2683, 2011.

[12] Z. Wang, H. R. Sheikh, and A.C. Bovik, “No-reference perceptual quality assessment of jpeg compressed images,” in *Image Processing, International Conference on*, 2002, vol. 1, pp. I–477–I–480 vol.1.

[13] Z Wang, L Lu, and A. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Comm.*, vol. 19, pp. 121–132, 2004.

[14] M H Pinson and S Wolf, “A new standardized method for objectively measuring video quality,” *Broadcasting, IEEE Transactions on*, vol. 50, no. 3, pp. 312–322, 2004.

[15] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ”completely blind” image quality analyzer.,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[16] A.K. Moorthy and A.C. Bovik, “A two-step framework for constructing blind image quality indices,” *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 513–516, May 2010.

[17] A. Mittal, A.K. Moorthy, and A.C. Bovik, “No-reference image quality assessment in the spatial domain,” *Image Processing, IEEE Trans. on*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.

[18] M.H. Pinson, C. Schmidmer, L. Janowski, R. Pepion, Quan H., P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, “Subjective and objective evaluation of an audiovisual subjective dataset for research and development,” in *Quality of Multimedia Experience, International Workshop on*, July 2013, pp. 30–31.