

Recover Subjective Quality Scores from Noisy Measurements

Zhi Li* and Christos G. Bampis†

*Netflix
100 Winchester Circle
Los Gatos, CA 95032, USA
zli@netflix.com

†Department of ECE
University of Texas at Austin
Austin, TX 78712, USA
bampis@utexas.edu

Abstract

Simple quality metrics such as PSNR are known to not correlate well with subjective quality when tested across a wide spectrum of video content or quality regime. Recently, efforts have been made in designing objective quality metrics trained on subjective data, demonstrating better correlation with video quality perceived by human. Clearly, the accuracy of such a metric heavily depends on the quality of the subjective data that it is trained on. In this paper, we propose a new approach to recover subjective quality scores from noisy raw measurements, by jointly estimating the subjective quality of impaired videos, the bias and consistency of test subjects, and the ambiguity of video contents all together. Compared to previous methods which partially exploit the subjective information, our approach is able to exploit the information in full, yielding better handling of outliers without the need for z-scoring or subject rejection. It also handles missing data more gracefully. Lastly, as side information, it provides interesting insights on the test subjects and video contents.

1 Introduction

In video coding research and development, two methods have generally been used to evaluate the quality of impaired videos: subjective assessment through viewer experiments and objective assessment using quality metrics. Subjective assessment is the ultimate measure of viewer’s perception of quality, but is usually expensive to conduct. Very often, objective assessment is used as an alternative or complement to report perceptual quality. Peak-signal-to-noise-ratio (PSNR) and Structural Similarity Index (SSIM) [1] are examples of objective quality metrics originally designed for images but later extended to video. Besides, an objective quality metric can also be used as an optimization objective function, such as in per-title encode optimization [2] and per-scene encode optimization under bitrate and video buffer constraints.

Simple metrics such as PSNR find success in evaluating small differences and close performance among codecs or coding tools. However, it is well known that they do not correlate well with quality perceived by human when tested across a wide spectrum of quality or content. Recently, efforts have been made in addressing this issue by designing objective quality metrics through *subjective data fusion* [3–5]. The basic idea is to extract low-level features or elementary metrics that are quality-indicative, and then use a machine-learning regressor, such as a support vector machine (SVM) [6] or a neural net, to fuse them into a “meta-metric” that makes a final prediction. The regressor model is trained using subjective data, such as mean opinion scores (MOS) aggregated over the raw opinion scores collected from subjective experiments. It is

shown that the fusion-based approach correlates better with subjective data than other approaches [3].

Clearly, the accuracy of a fusion-based metric heavily depends on the quality of the subjective data that it is trained on. Thus, it is very important to provide clean and reliable training data. On the other hand, raw opinion scores offered by viewers are often noisy and unreliable, due to the following reasons:

- *Subject bias.* The notion of quality is highly subjective and test subjects are entitled to rate the videos in their own opinions. For example, more picky viewers tend to be biased toward lower scores, and vice versa. Also, not every subject has “golden eyes” – their sensitivity to impairments varies.
- *Subject inconsistency.* Subjective testing is a laborious process, and not every viewer can maintain attentiveness throughout. Some tend to rate more consistently than others.
- *Content ambiguity.* Some contents tend to be more difficult to be rated than others. For example, water surface with ripples in the dark is more ambiguous than a bright blue sky.
- *Outliers.* Last but not least, some raw scores are simply outliers – viewers may just not pay attention. Software issues may also render scores meaningless.¹

Existing approaches address some of the issues above. For example, MOS averages the raw scores from a number of subjects to produce an aggregate score, compensating for the bias and inconsistency of individuals. Z-score transformation (or z-scoring) [7] normalizes the scores on a per-subject basis. Subject rejection [8] counts the number of instances when a subject’s rating looks like an outlier, and if this occurs too often, the subject and all his or her scores are rejected.

In this paper, we propose a new approach to recover subjective quality scores from the noisy raw opinion measurements, by *jointly estimating* the subjective quality of impaired videos, the bias and consistency of test subjects, and the ambiguity of video contents all together. We propose a generative model that incorporates random variables representing each of these factors, cast the problem as maximum likelihood estimation (MLE), and derive a solution based on belief propagation (BP) [9]. Compared to previous methods which partially exploit the subjective information, our approach is able to exploit the information in full, yielding better handling of outliers without the need for z-scoring or subject rejection. It also handles missing data more gracefully. Lastly, as side information, it provides interesting insights on the test subjects and video contents.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 defines the problem and notations. Section 4 describes traditional approaches, and Section 5 presents the proposed approach. Experimental results are reported in Section 6.

This work’s open-source implementation can be found at [10].

¹We have seen real examples where some raw scores get misaligned with subjects and contents due to a software bug!

2 Related Work

The ITU-R BT.500 Recommendation [8] defines procedures for video subjective testings including both single-stimulus and double-stimulus methods. It also defines the method for subjective rejection. The ITU-T P.910 Recommendation [11] defines procedures for calculating differential MOS. Z-score transformation of subjective scores is proposed in [7] as a pre-processing step prior to subject rejection.

A theoretical model for subjects' influence on test scores is proposed in [12], which is similar in spirit to ours. The authors have focused on validating the model using real data, which also provides support for this work. However, they have relied on repetitive experiments for model parameter estimation, while our proposed solution based on BP is much more efficient.

Recovering subjective quality scores from noisy measurements is closely related to the task of label inference from very large databases of hand labeled images. To address the label inference problem, [13] proposes a solution based on probabilistic graphical model. [14] further extends the idea to the task of image quality evaluation. There is a number of differences between their approach and this work. First, they formulate a classification problem whereas this work considers a regression problem, which is closer to the nature of subjective test procedure considered. Second, their work adopts a discriminative model whereas this work adopts a generative model, allowing better interpretability of results.

3 Problem Formulation

Consider an experiment with S subjects, indexed by $s = 1, \dots, S$, and E impaired video encodes, indexed by $e = 1, \dots, E$. For simplicity, we consider the case of a single viewing session with no repetitions. A subject s rates an impaired video encode e , producing a raw opinion score $x_{e,s}$. In the *full sampling* scenario, every subject rates every impaired video. In the *selective sampling* scenario, not every subject needs to rate every impaired video – if a score is missing, it is denoted by $x_{e,s} = *$. In the rest of the paper, unless otherwise stated, the full sampling scenario is assumed.

The raw opinion scores may be produced using different test methods, including both single-stimulus and double-stimulus methods: In *absolute category rating (ACR)*, the subject is instructed to watch the impaired video and give a rating on the scale from 1 (quality is bad) to 5 (quality is excellent). In *degradation category rating (DCR)*, the subject is instructed to watch a pair of videos – the unimpaired reference video followed by the impaired video, and then rate the impaired video on the scale from 1 (impairments are very annoying) to 5 (impairments are imperceptible). In either method, unimpaired *hidden reference* videos may be present along with other impaired videos, and are also rated by the subject. A differential score between the score of the impaired and its hidden reference may be used in place of the raw opinion score, and the resulting MOS calculated is called differential MOS (or DMOS) [11]. DMOS is useful when the quality degradation from the reference is relevant, rather than the absolute quality when, for example, the reference video is not perfect in quality.

Algorithm 1 Subject rejection [8]

- Input: $x_{e,s}$ for $s = 1, \dots, S$ and $e = 1, \dots, E$.
 - Initialize $p(s) \leftarrow 0$ and $q(s) \leftarrow 0$ for $s = 1, \dots, S$.
 - For $e = 1, \dots, E$:
 - Let $Kurtosis_e = \frac{m_{4,e}}{m_{2,e}^2}$.
 - If $2 \leq Kurtosis_e \leq 4$, then $\epsilon_e = 2$; otherwise $\epsilon_e = \sqrt{20}$.
 - For $s = 1, \dots, S$:
 - * If $x_{e,s} \geq \mu_e + \epsilon_e \sigma_e$, then $p(s) \leftarrow p(s) + 1$.
 - * If $x_{e,s} \leq \mu_e - \epsilon_e \sigma_e$, then $q(s) \leftarrow q(s) + 1$.
 - Initialize $Set_{rej} = \emptyset$.
 - For $s = 1, \dots, S$:
 - If $\frac{p(s)+q(s)}{E} \geq 0.05$ and $\left| \frac{p(s)-q(s)}{p(s)+q(s)} \right| < 0.3$, then $Set_{rej} \leftarrow Set_{rej} \cup \{s\}$.
 - Output: Set_{rej} .
-

Each impaired video e is also associated with a *content* c , denoted by $c = \mathbf{c}(e)$, with $c = 1, \dots, C$ where C is the total number of video contents used in the experiment.

Throughout the paper, we use μ_e and μ_s to denote the mean values calculated over scores for impaired video e and for subject s , respectively, i.e., $\mu_e = \frac{1}{S} \sum_s x_{e,s}$ and $\mu_s = \frac{1}{E} \sum_e x_{e,s}$. Similarly, $m_{n,e}$ and $m_{n,s}$ denote the n -th order central moment over scores for e and s , respectively, i.e., $m_{n,e} = \frac{1}{S} \sum_s (x_{e,s} - \mu_e)^n$ and $m_{n,s} = \frac{1}{E} \sum_e (x_{e,s} - \mu_s)^n$. As special case, the standard deviation over scores for e and s have the form $\sigma_e = \sqrt{m_{2,e}}$ and $\sigma_s = \sqrt{m_{2,s}}$, respectively.

4 Traditional Approaches

The most basic approach to recover subjective quality scores from the raw opinion measurements is by averaging, or MOS: $x_e = \mu_e$, for impaired video $e = 1, \dots, E$. Before the averaging step, one can apply the following:

Z-score transformation. It is advocated that the scores undergo a normalization procedure on a per-subject basis [7], i.e., $x_{e,s} \leftarrow \frac{x_{e,s} - \mu_s}{\sigma_s}$. Applying z-scoring could compensate for individual subject's bias and inconsistency, preparing for favorable conditions for subject rejection. However, after z-scoring, the original scale of the scores is unfavorably lost, leading to difficulty in interpreting the results. Also, it only partially compensates for the influence of subjects. A theoretical analysis on z-scoring can be found in Appendix A.

Subject rejection. In [8], a recommendation for subject rejection is provided. The algorithm is reproduced in Algorithm 1 for completeness. Video by video, the algorithm counts the number of instances when a subject's opinion score deviates by a few sigmas, and reject the subject if the occurrences are more than a fraction. All scores corresponding to the rejected subjects are discarded, which could be an overkill.

5 Proposed Approach

In this section, we describe the proposed approach of jointly estimating the subjective quality of impaired videos, the bias and consistency of subjects, and the ambiguity of video contents. Different from traditional approaches which require an explicit z-score transformation or a subject rejection step, the proposed model naturally accounts for subjective biases, inconsistencies and outliers.

5.1 The Model

We model the raw opinion scores as a random variable $\{X_{e,s}\}$ with the following form:

$$\begin{aligned} X_{e,s} &= x_e + B_{e,s} + A_{e,s}, \\ B_{e,s} &\sim \mathcal{N}(b_s, v_s^2), \\ A_{e,s} &\sim \mathcal{N}(0, a_{c:c(e)=c}^2) \end{aligned} \quad (1)$$

for $e = 1, \dots, E$ and $s = 1, \dots, S$. In this model, x_e represents the quality of impaired video e perceived by an *average* viewer, $B_{e,s}$, $e = 1, \dots, E$ are i.i.d. Gaussian variables representing the factor of subject s , and $A_{e,s}$, $s = 1, \dots, S$ and $e : c(e) = c$ are i.i.d. Gaussian variables representing the factor of video content c (i.e., the content that e corresponds to). The parameters b_s and v_s^2 represent the bias (i.e., mean) and inconsistency (i.e., variance) of subject s , $s = 1, \dots, S$. The parameter a_c^2 represents the ambiguity (i.e., variance) of content c , $c = 1, \dots, C$. In this formulation, the unknowns are the model parameters $\theta = (\{x_e\}, \{b_s\}, \{v_s\}, \{a_c\})$, where $\{\cdot\}$ denotes the corresponding set. The main idea of our approach is to jointly recover these unknowns by MLE. Let $L = \log P(\{x_{e,s}\}|\theta)$ be the log likelihood function, the goal is to solve for $\theta^* = \arg \max_{\theta} L$. Our ultimate goal is to recover the estimated scores $\{x_e\}$ of a hypothetical unbiased and consistent viewer, while $\{b_s\}$, $\{v_s\}$ and $\{a_c\}$ are side information about the subjects and video contents. An analysis on the parameters' recoverability can be found in Appendix B.

5.2 BP Solution

We derive a solution for the MLE formulation using BP algorithm. We start with the log-likelihood function. By (1), $X_{e,s}$ is a sum of a constant and independent Gaussian variables, thus $X_{e,s}$ is also Gaussian with $X_{e,s} \sim \mathcal{N}(x_e + b_s, v_s^2 + a_{c(e)}^2)$. The log-likelihood function can be expressed as:

$$\begin{aligned} L(\theta) &= \log P(\{x_{e,s}\}|\theta) \\ &= \log P(\{x_{e,s}\}|\{x_e\}, \{b_s\}, \{v_s\}, \{a_c\}) \\ &= \log \prod_{e,s} P(x_{e,s}|x_e, b_s, v_s, a_{c(e)}) \end{aligned} \quad (2)$$

$$\begin{aligned} &= \sum_{e,s} \log P(x_{e,s}|x_e, b_s, v_s, a_{c(e)}) \\ &\equiv \sum_{e,s} -\frac{1}{2} \log(v_s^2 + a_{c(e)}^2) - \frac{1}{2} \cdot \frac{(x_{e,s} - x_e - b_s)^2}{v_s^2 + a_{c(e)}^2} \end{aligned} \quad (3)$$

Algorithm 2 BP solution for the proposed MLE formulation

- Input:
 - $x_{e,s}$ for $s = 1, \dots, S$ and $e = 1, \dots, E$.
 - Refresh rate α .
 - Stop threshold Δx^{thr} .
 - Initialize $\{x_e\} \leftarrow \{\mu_e\}$, $\{b_s\} \leftarrow \{0\}$, $\{v_s\} \leftarrow \{\sigma_s\}$, $\{a_c\} \leftarrow \{\sigma_c\}$.
 - Loop:
 - $\{x_e^{prev}\} \leftarrow \{x_e\}$.
 - $b_s \leftarrow (1 - \alpha) \cdot b_s + \alpha \cdot b_s^{new}$ where $b_s^{new} = b_s - \frac{\partial L(\theta)/\partial b_s}{\partial^2 L(\theta)/\partial b_s^2}$ for $s = 1, \dots, S$.
 - $v_s \leftarrow (1 - \alpha) \cdot v_s + \alpha \cdot v_s^{new}$ where $v_s^{new} = v_s - \frac{\partial L(\theta)/\partial v_s}{\partial^2 L(\theta)/\partial v_s^2}$ for $s = 1, \dots, S$.
 - $a_c \leftarrow (1 - \alpha) \cdot a_c + \alpha \cdot a_c^{new}$ where $a_c^{new} = a_c - \frac{\partial L(\theta)/\partial a_c}{\partial^2 L(\theta)/\partial a_c^2}$ for $c = 1, \dots, C$.
 - $x_e \leftarrow (1 - \alpha) \cdot x_e + \alpha \cdot x_e^{new}$ where $x_e^{new} = x_e - \frac{\partial L(\theta)/\partial x_e}{\partial^2 L(\theta)/\partial x_e^2}$ for $e = 1, \dots, E$.
 - If $\left(\sum_{e=1}^E (x_e - x_e^{prev})^2\right)^{\frac{1}{2}} < \Delta x^{thr}$, break.
 - Output: $\{x_e\}$, $\{b_s\}$, $\{v_s\}$, $\{a_c\}$.
-

where (2) uses the independence assumption on opinion scores and (3) uses the Gaussian formula with omission of the constant terms. With (3), the first- and second-order partial derivatives of $L(\theta)$ with respect to parameters x_e , b_s , v_s and a_c can be derived. We then apply the Newton-Raphson rule [9] $v \leftarrow v - \frac{\partial L/\partial v}{\partial^2 L/\partial v^2}$ to update each parameter v at a time in each iteration. Note that other update rules are also possible, but using the Newton-Raphson rule can yield nice expressions with interpretability. Also note that the BP algorithm finds a local optimal solution when the problem is nonconvex. It is important to initialize the parameters properly. We choose the MOS as the initial values for $\{x_e\}$, zeros for $\{b_s\}$, and the standard deviation values $\{\sigma_s\}$ and $\{\sigma_c\}$ for $\{v_s\}$ and $\{a_c\}$, respectively, where $\sigma_c^2 = \frac{\sum_{s,e:c(e)=c} (x_{e,s} - \mu_s)^2}{\sum_{s,e:c(e)=c} 1}$. The BP algorithm solution for the proposed MLE formulation is summarized in Algorithm 2. The analytical forms of the update rules are derived in Appendix C. A good choice of refresh rate and stop threshold are $\alpha = 0.1$ and $\Delta x^{thr} = 1e^{-9}$, respectively.

5.3 Generalization to Selective Sampling

All the algorithms described so far, including the traditional approaches and the proposed MLE formulation, assumed the full sampling scenario. It is not difficult to generalize the algorithms to selective sampling. Simply exclude the missing terms during summation, i.e., use $\sum_{e:x_{e,s} \neq *} f(x_{e,s})$ instead of $\sum_e f(x_{e,s})$, for some function $f(\cdot)$. The mean and central moment terms now become: $\mu_e = \frac{\sum_{s:x_{e,s} \neq *} x_{e,s}}{\sum_{s:x_{e,s} \neq *} 1}$, $\mu_s = \frac{\sum_{e:x_{e,s} \neq *} x_{e,s}}{\sum_{e:x_{e,s} \neq *} 1}$, $m_{n,e} = \frac{\sum_{s:x_{e,s} \neq *} (x_{e,s} - \mu_e)^n}{\sum_{s:x_{e,s} \neq *} 1}$ and $m_{n,s} = \frac{\sum_{e:x_{e,s} \neq *} (x_{e,s} - \mu_s)^n}{\sum_{e:x_{e,s} \neq *} 1}$.

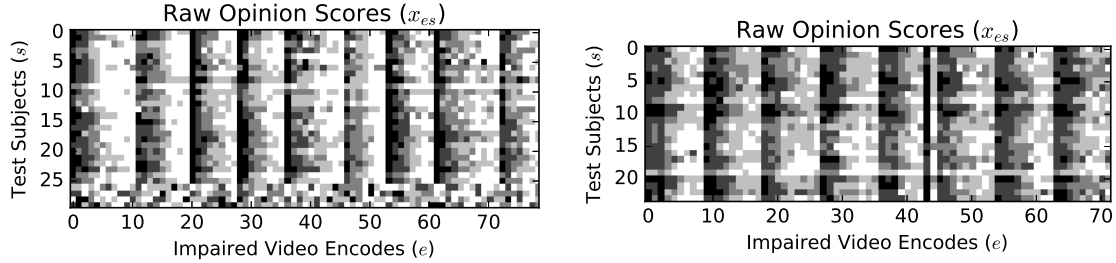


Figure 1: Raw opinion scores from (left) the NFLX dataset and (right) the VQEG dataset. Each pixel represents a raw opinion score. The darker the color, the lower the score. The impaired videos are arranged by contents, and within each content, from low quality to high quality (with the reference video always in the last). For the NFLX dataset, the last four rows correspond to corrupted subjective data.

6 Results

We set up a number of experiments to evaluate the performance of the proposed MLE method and compare it with a number of traditional methods, including the plain MOS, MOS with subject rejection (SR-MOS) and MOS with z-scoring and subject rejection (ZS-SR-MOS).

We use raw opinion scores from two datasets: the Netflix Public (NFLX) dataset [15] and the VQEG HD3 (VQEG) dataset [16]. Refer to Figure 1 for a visualization of the raw scores. The NFLX dataset includes four subjects whose raw scores were scrambled due to a software issue during data collection. The VQEG dataset includes various contents (SRC01-09 excluding 04 which overlaps with the NFLX dataset) and streaming-relevant impairments (HRC04, 07 and 16-21). Note that the SRC06-HRC07 video received very low scores due to encoding issues.

For many of the experiments, we do not have “ground truth” quality scores to compare against (we only have noisy raw scores). Instead, we use the following methodology in our report of results. For each recovery method, we have a benchmark result, which is the recovered quality scores obtained *using that method* (for fairness) on an *unaltered full dataset*. The quality scores recovered under certain conditions (e.g., using a portion of the raw scores, partially corrupted) is compared against the benchmark, and a root-mean-squared-error (RMSE) value is reported. In doing so, we could evaluate, for example, how fast the results converge toward that benchmark result. In other experiments, we do have a “ground truth”, when artificially omitting a subject or creating a corruption on the data.

6.1 An Example

Let us first visually inspect one example result recovered by the MLE and compare it with the MOS. Figure 2 shows the recovered parameters on the full NFLX dataset. Comparing the result with Figure 1 (left), there are a few observations: 1) The quality scores recovered by MLE are numerically different from the MOS, suggesting that the recovery is non-trivial. 2) Subject #10 has the highest bias, which is evidenced by the whitish horizontal strip visible in Figure 1 (left). 3) The last four subjects (whose

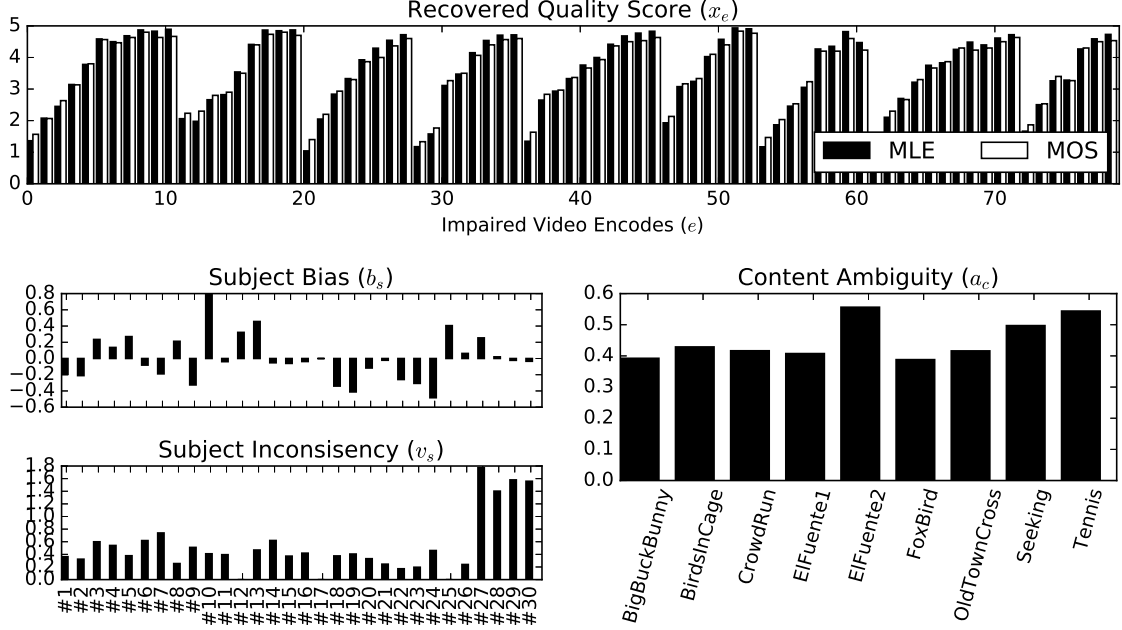


Figure 2: Example recovery result on the NFLX dataset.

raw scores were scrambled) have a very high a_c value. 4) The content with the highest ambiguity is ElFuente2, which is the “fountain and toddler” scene, known to be difficult to evaluate. These observations demonstrate the potential of the MLE method on the problem at hand.

6.2 Convergence

Next, we evaluate how fast MLE and other methods converge toward the result recovered on the full dataset, as we increase the number of test subjects. For each number of subjects, we randomly sample among all the subjects, and repeated the experiment 100 times. Figure 3 illustrates the averaged RMSE as a function of the subject numbers. It is shown that on the NFLX dataset, MLE has the closest quality scores to the full-dataset recovery than other methods for a given number of subjects. On the VQEG dataset, MLE has performance comparable to MOS and SR-MOS.

6.3 Resistance to Corruption

From the last section, it is evident that MLE shows faster convergence than other methods in the presence of data scrambling. To further corroborate our speculation, we evaluate how MLE and other methods behave in the presence of data corruption. For each dataset, we simulate two cases of corruption: a) subject corruption, where all the scores corresponding to a number of subjects are scrambled, and b) random corruption, where a raw score gets replaced by a random score from 1 to 5 with a probability. Results for a) and b) are reported in Figure 4 and 5, respectively.

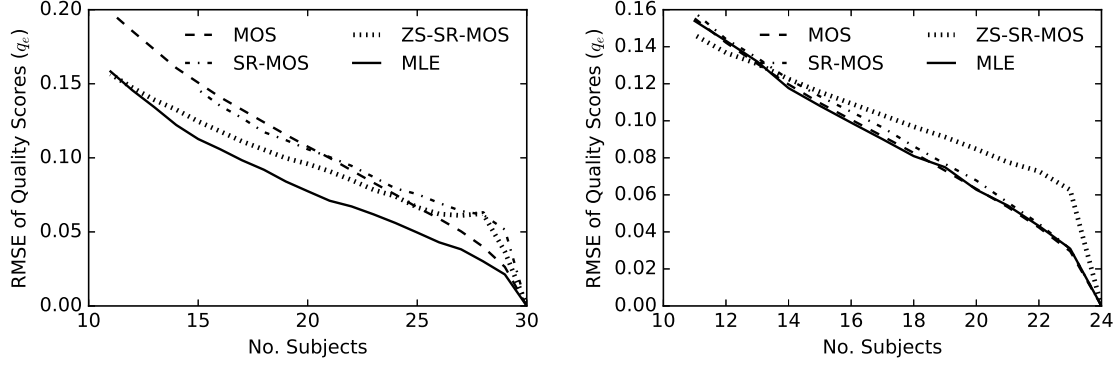


Figure 3: RMSE of the recovered subjective quality scores as a function of the number of test subjects, on (left) the NFLX dataset and (right) the VQEG dataset.

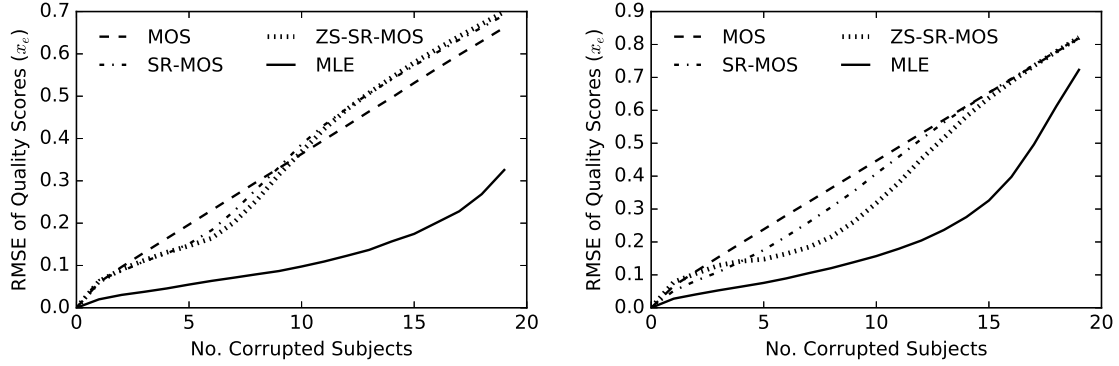


Figure 4: RMSE of the recovered subjective quality scores as a function of the number of corrupted test subjects, on (left) the NFLX dataset and (right) the VQEG dataset.

It can be observed that in the presence of subject corruption, MLE achieves a substantial gain over the other methods, including the ones with subject rejection. The reason is that the proposed model was able to capture the variance of subjects explicitly and is able to compensate for it. On the other hand, the traditional subject rejection scheme (Algorithm 1) was only able to identify part of the corrupted subjects. It may also occur that only a subset of a subject's scores is unreliable. In that case, discarding all of the subject's scores is a waste of valuable subjective data. Meanwhile, traditional subject rejection employs a set of heuristic steps to determine outliers, which may lack interpretability. By contrast, the proposed model naturally integrates the various subjective effects together and is solved efficiently by our MLE method.

In the presence of random corruption, it can be seen that MLE does not show any advantage over the other methods. This is because the proposed model (1) is incapable of capturing this type of corruption, hence it could not deal with it effectively.

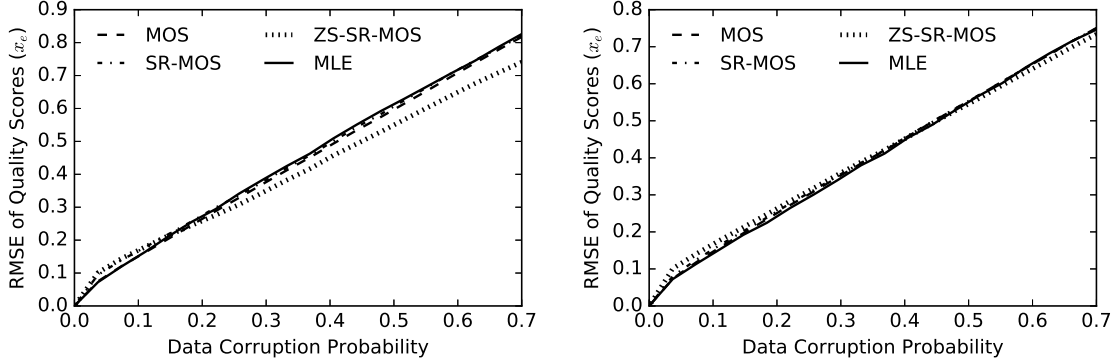


Figure 5: RMSE of the recovered subjective quality scores as a function of the random score corruption probability, on (left) the NFLX dataset and (right) the VQEG dataset.

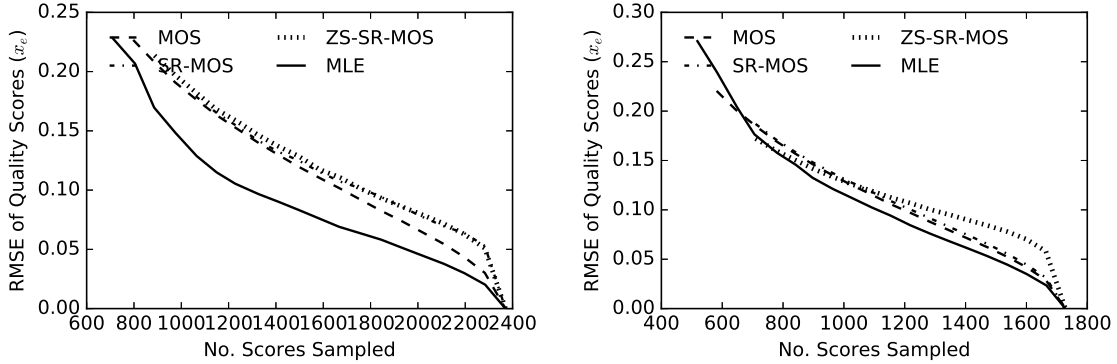


Figure 6: RMSE of the recovered subjective quality scores as a function of the number of scores randomly sampled, on (left) the NFLX dataset and (right) the VQEG dataset.

6.4 Selective Sampling

We also evaluate the MLE and other methods under selective sampling, where each subject only rates a part of all videos. We performed this step by assigning each raw score a random probability of presence. As the probability increases, more scores are sampled. The performance is reported in Figure 6. On the NFLX dataset, again, MLE has clear advantage over other methods. On the VQEG dataset, MLE also shows gain over other methods. Since MLE accounts for the full information, randomly missing some data points does not affect its predictive performance by much. By contrast, the MOS, SR-MOS, ZS-SR-MOS methods, which make local decisions on partial information, are greatly affected.

7 Summary and Future Work

We have presented a new approach to process raw opinion scores collected in video quality subjective testing, by using a generative model that jointly captures the quality of impaired videos, bias and inconsistency of subjects and the ambiguity of contents. We determined the model parameters by formulating a MLE problem and devising a BP solution. It was shown that the recovered parameters were able to

capture the subjective effects (bias, inconsistencies etc.) of the problem at hand, and that the proposed solution outperformed other methods in terms of its resistance to subject corruption, better handling of missing data and provision of side information on the test subjects and video contents.

We list a number of directions to further this work: 1) The recovered side information on the test subjects and video contents can be used actively in the process to train an objective quality metric to yield more accurate prediction. 2) The proposed model can be further extended to other test methods, such as pairwise comparison. 3) A theoretical analysis on the BP algorithm can uncover convergence region of its parameters.

References

- [1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.
- [2] A. Aaron, Z. Li, M. Manohara, J. D. Cock, and D. Ronca. Per-Title Encode Optimization. [Online]. Available: <http://techblog.netflix.com/2015/12/per-title-encode-optimization.html>
- [3] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward A Practical Perceptual Video Quality Metric. [Online]. Available: <http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html>
- [4] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A Fusion-based Video Quality Assessment (FVQA) Index," *APSIPA Trans. Signal and Information Processing*, 2014.
- [5] S. Wolf and M. H. Pinson, "Video Quality Model for Variable Frame Delay (VQM-VFD)," Nat. Telecommun. Inf. Admin., Tech. Memo, TM-11-482, Sep. 2011.
- [6] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [7] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.
- [8] ITU-R BT.500: Methodology for the Subjective Assessment of the Quality of Television Pictures. [Online]. Available: <https://www.itu.int/rec/R-REC-BT.500>
- [9] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [10] VMAF Open-Source Project. [Online]. Available: <https://github.com/Netflix/vmaf>
- [11] ITU-T P.910: Subjective Video Quality Assessment Methods for Multimedia Applications. [Online]. Available: <https://www.itu.int/rec/T-REC-P.910>
- [12] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec 2015.
- [13] J. Whitehill, T. fan Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems*, 2009, pp. 2035–2043.
- [14] W. Wang, J. Allebach, and Y. Guo, "Image quality evaluation using image quality ruler and graphical model," in *IEEE Int. Conf. Image Processing (ICIP)*, 2015.
- [15] Netflix Public Dataset. [Online]. Available: <https://github.com/Netflix/vmaf#netflix-public-dataset>

- [16] “Report on the Validation of Video Quality Models for High Definition Video Content,”
Video Quality Experts Group (VQEG), Tech. Rep., Version 2.0, June 30, 2010.

A Analysis on Z-score Transformation

Let us assume a simplified version of model (1):

$$\begin{aligned} X_{e,s} &= x_e + B_{e,s} \\ B_{e,s} &\sim \mathcal{N}(b_s, v_s^2). \end{aligned}$$

Let $y_{e,s} = x_{e,s} - \mu_s$ be the mean-subtracted version of $x_{e,s}$, and let $B_{e,s} = b_s + v_s N_{e,s}$ where $N_{e,s} \sim \mathcal{N}(0, 1)$. The corresponding random variable $Y_{e,s}$ has the form:

$$\begin{aligned} Y_{e,s} &= X_{e,s} - \frac{1}{E} \sum_{e'} X_{e',s} \\ &= x_e + b_s + v_s N_{e,s} - \frac{1}{E} \sum_{e'} (x_{e'} + b_s + v_s N_{e',s}) \\ &= (x_e - \frac{1}{E} \sum_{e'} x_{e'}) + v_s N_{e,s} - \frac{1}{E} v_s \sum_{e'} N_{e',s} \\ &= y_e + \frac{E-1}{E} v_s N_{e,s} - \frac{1}{E} v_s \sum_{e': e' \neq e} N_{e',s} \end{aligned} \tag{4}$$

$$\simeq y_e + v_s N_{e,s}, \tag{5}$$

where in (4) we define $y_e = x_e - \mu_x$ and $\mu_x = \frac{1}{E} \sum_{e'} x_{e'}$, and (5) is due to $E \gg 1$. Z-score transformation computes the z-score $z_{e,s} = (x_{e,s} - \mu_s) / \sigma_s = y_{e,s} / \sqrt{\frac{1}{E} \sum_{e'} y_{e',s}^2}$. The corresponding random variable $Z_{e,s}$ has the form:

$$\begin{aligned} Z_{e,s} &= \frac{Y_{e,s}}{\sqrt{\frac{1}{E} \sum_{e'} Y_{e',s}^2}} \\ &\simeq \frac{y_e + v_s N_{e,s}}{\sqrt{\frac{1}{E} \sum_{e'} (y_{e'} + v_s N_{e',s})^2}} \\ &\simeq \frac{y_e + v_s N_{e,s}}{\sqrt{\frac{1}{E} \sum_{e'} y_{e'}^2 + \frac{1}{E} v_s^2 \sum_{e'} N_{e',s}^2}} \end{aligned} \tag{6}$$

$$\begin{aligned} &= \frac{x_e - \mu_x + v_s N_{e,s}}{\sqrt{\sigma_x^2 + \frac{1}{E} v_s^2 \sum_{e'} N_{e',s}^2}} \\ &\neq \frac{x_e - \mu_x}{\sigma_x}, \end{aligned} \tag{7}$$

where (6) is due to that $\sum_e y_e = 0$ and in (7) we define $\sigma_x = \sqrt{\frac{1}{E} \sum_e y_e^2}$. The last inequality suggests that there are non-vanishing terms related to subjects that cannot be canceled out by z-score transformation. In other words, z-scoring can only partially compensate subjects' bias and inconsistency.

B Parameter Recoverability by MLE

We consider the recoverability of parameters in (1) by MLE. Assume a typical scenario, where the subjective experiment has $S = 30$ subjects, $C = 20$ contents, and $E = 200$ impaired videos. The number of unknowns is thus $E + 2S + C = 280$ and the number of observations is $ES = 6000$. At a first glance, we may conclude that the number of observations are much greater than the number of unknowns, and the problem has a well-conditioned solution.

Sounds pretty good? Let's take a closer look. This time, consider a simplified model by removing the variation terms $\{v_s\}$ and $\{a_c\}$, i.e.,

$$x_{e,s} = x_e + b_s.$$

After the variation terms are removed, everything is deterministic. Since the relationship is linear, we hope to find a least-squares solution of a linear system with ES observations and $E + S$ unknowns. Let $u = [\{x_e\}_{e=1}^E, \{b_s\}_{s=1}^S]^T$ be the vector of unknowns, and $o = [\{x_{e,s}\}]^T$ be the vector of observations, and let $\Omega = [\Omega_{i,j}]$ be a $ES \times (E + S)$ matrix, where $\Omega_{e+(s-1)E,e} = 1$ and $\Omega_{e+(s-1)E,s+E} = 1$ for $e = 1, \dots, E$ and $s = 1, \dots, S$, and the rest of the elements are all zeros. We look to find a least-squares solution to the following linear system:

$$\Omega \cdot u = o.$$

Proposition 1. *Matrix Ω has rank $E + S - 1$.*

Proof. Apply Gaussian elimination. The first E rows are independent. Subtract the $(e + (s - 1)E)$ -th row by the e -th row, $e = 1, \dots, E$ and $s = 1, \dots, S$, we can get another $(S - 1)$ independent rows. In total, there are $E + S - 1$ independent rows in Ω . \square

Proposition 1 implies that we are missing one equation, and without it there are infinite number of solutions. Fortunately, it is not too difficult to come up with one additional equation that is reasonable. For example, we can let the first impaired video have a score equal to its MOS, i.e.,

$$x_{e=1} = \mu_{e=1}.$$

We can readily find the least-squares solution after adding the additional condition, which is also a necessary condition for solving the full model (1).

C Derivation of Update Rules in Algorithm 2

This section derives the update rules used in Algorithm 2. Consider $L(\theta)$ derived in (3) where $\theta = (\{x_e\}, \{b_s\}, \{v_s\}, \{a_c\})$. The first-order partial derivatives of $L(\theta)$ are expressed as follows:

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial x_e} &= \sum_s \frac{x_{e,s} - x_e - b_s}{v_s^2 + a_{c(e)}^2} \\
\frac{\partial L(\theta)}{\partial b_s} &= \sum_e \frac{x_{e,s} - x_e - b_s}{v_s^2 + a_{c(e)}^2} \\
\frac{\partial L(\theta)}{\partial v_s} &= \sum_e -\frac{v_s}{v_s^2 + a_{c(e)}^2} + \frac{v_s(x_{e,s} - x_e - b_s)^2}{(v_s^2 + a_{c(e)}^2)^2} \\
\frac{\partial L(\theta)}{\partial a_c} &= \sum_{s,e:c(e)=c} -\frac{a_{c(e)}}{v_s^2 + a_{c(e)}^2} + \frac{a_{c(e)}(x_{e,s} - x_e - b_s)^2}{(v_s^2 + a_{c(e)}^2)^2}
\end{aligned}$$

Repeat the differentiation and we get the second-order partial derivatives:

$$\begin{aligned}
\frac{\partial^2 L(\theta)}{\partial x_e^2} &= \sum_s -\frac{1}{v_s^2 + a_{c(e)}^2} \\
\frac{\partial^2 L(\theta)}{\partial b_s^2} &= \sum_e -\frac{1}{v_s^2 + a_{c(e)}^2} \\
\frac{\partial^2 L(\theta)}{\partial v_s^2} &= \sum_e -\frac{a_{c(e)}^2 - v_s^2}{(v_s^2 + a_{c(e)}^2)^2} + \frac{(x_{e,s} - x_e - b_s)^2 (a_{c(e)}^4 - 2a_{c(e)}^2 v_s^2 - 3v_s^4)}{(v_s^2 + a_{c(e)}^2)^4} \\
\frac{\partial^2 L(\theta)}{\partial a_c^2} &= \sum_{s,e:c(e)=c} -\frac{v_s^2 - a_{c(e)}^2}{(v_s^2 + a_{c(e)}^2)^2} + \frac{(x_{e,s} - x_e - b_s)^2 (v_s^4 - 2v_s^2 a_{c(e)}^2 - 3a_{c(e)}^4)}{(v_s^2 + a_{c(e)}^2)^4}
\end{aligned}$$

Define $w_{e,s} = \frac{1}{v_s^2 + a_{c(e)}^2}$. Applying the expressions above to the Newton-Raphson update rules yield:

$$\begin{aligned}
x_e^{new} &= \frac{\sum_s w_{e,s}(x_{e,s} - b_s)}{\sum_s w_{e,s}} \\
b_s^{new} &= \frac{\sum_e w_{e,s}(x_{e,s} - x_e)}{\sum_e w_{e,s}} \\
v_s^{new} &= v_s - \frac{\sum_e w_{e,s} v_s - w_{e,s}^2 v_s (x_{e,s} - x_e - b_s)^2}{\sum_e w_{e,s}^2 (a_{c(e)}^2 - v_s^2) - w_{e,s}^4 (x_{e,s} - x_e - b_s)^2 (a_{c(e)}^4 - 2a_{c(e)}^2 v_s^2 - 3v_s^4)} \\
a_c^{new} &= a_c - \frac{\sum_{s,e:c(e)=c} w_{e,s} a_{c(e)} - w_{e,s}^2 a_{c(e)} (x_{e,s} - x_e - b_s)^2}{\sum_{s,e:c(e)=c} w_{e,s}^2 (v_s^2 - a_{c(e)}^2) - w_{e,s}^4 (x_{e,s} - x_e - b_s)^2 (v_s^4 - 2v_s^2 a_{c(e)}^2 - 3a_{c(e)}^4)}
\end{aligned}$$

Note that there is strong intuition behind the expressions for x_e^{new} and b_s^{new} . In each iteration, x_e^{new} is re-estimated, by the sum of opinion scores $x_{e,s}$ with the currently

estimated bias b_s removed. Each opinion score is weighted by $w_{e,s}$, i.e., the higher the variance (subject inconsistency and content ambiguity), the less reliable the opinion score, hence less the weight. We can interpret b_s^{new} in a similar way.