ZREC: ROBUST RECOVERY OF MEAN AND PERCENTILE OPINION SCORES

Jingwen Zhu[1§], Ali Ak[1§], Patrick Le Callet[1], Sriram Sethuraman[2], Kumar Rahul[2]

[1]Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
[2]Amazon Prime Video, Bangalore, India

ABSTRACT

Observer screening and subject opinion score recovery is essential for collecting a reliable QoE database. This paper proposes a new method, ZREC*, which uses Z-scores to estimate subject bias, inconsistency, and content ambiguity. Additionally, we propose Mean Opinion Score (MOS) recovery and Percentile Opinion Score (POS) recovery scheme based on the three estimated parameters. ZREC does not fully reject subjects, rather adjust their coefficients in the MOS/POS recovery, allowing for more efficient use of data collection. The estimated parameters of ZREC are highly correlated with more complex solver-based methods and standards. In addition, ZREC recovers MOS with smaller confidence intervals than the state of the art. Experimental results also demonstrate that using recovered p_{th} POS as ground truth during training improves the performance of Satisfied User Ratio (SUR) prediction.

***Index Terms*—** Observer screening, MOS recovery, Satisfied User Ratio, Quality of Experience

1. INTRODUCTION

Observer screening is essential steps of collecting a reliable Quality of Experience (QoE) database. A range of methods [1, 2, 3, 4, 5] with varying complexities have been proposed, and various standards [6, 7, 8] include recommendations for this purpose. It has also been demonstrated that training learning-based metrics on recovered MOS can enhance their performance to a certain extent [9].

The collected QoE measurements in subjective studies are often characterized as a combination of subject bias, inconsistency and the underlying quality of the stimuli [1]. **Subject bias** refers to the systematic error of a subject towards a certain direction, *e.g.* a positive bias indicates the subjects overall tendency to perceive a higher quality. **Subject inconsistency** is associated with the random unexplained error included in the observations, such as lack of attention, malicious intentions etc. On another front, the **content ambiguity** defines the level of difficulty in evaluating a stimulus due to its inherent ambiguity.

§Equal contribution
*available at: https://github.com/kyillene/ZREC

Table 1. Summary of the estimated parameters by each mos recovery model.

	BT500	P913-12.4	P913-12.6	MLE	ZREC
Subject Inconsistency	✗	✗	✓	✓	✓
Subject Bias	✗	✓	✓	✓	✓
Content Ambiguity	✗	✗	✗	✓	✓

Below, we provide a summary of commonly used MOS recovery methods from the literature and briefly discuss their advantages and disadvantages. In addition, Table 1 provides a quick overview of the parameters estimated by each model.

- **BT500:** ITU-R BT.500 Recommendation [6] defines an **outlier rejection** procedure. Subjects are rejected based on the number of opinion scores outside of the predefined amount of standard deviation range of the population. If a subject found to be an outlier, all of his/her opinions are removed from the dataset. MOS is calculated as the mean of remaining subjects. Due to hard-coded thresholds and removing all votes of a detected outlier, the MOS recovery may result in even larger confidence intervals
- **P913-12.4:** ITU-R P.913 Recommendation clause 12.4 [8] proposes a procedure based on both **bias removal** and **outlier rejection**. For each subject, the bias is calculated as the average difference between the MOS and subjects' opinion score of each stimulus. Estimated biases are removed from subject opinion scores and then MOS can be calculated as the average of bias-removed opinion scores, optionally after rejecting outliers. It can be seen that P913-12.4 is a slight improvement over BT500. However after removing the subject bias, the observer are still rejected with hard-coded parameters or treated equally by ignoring the subject inconsistency.
- **P913-12.6:** ITU-R P.913 Recommendation clause 12.6 [8] defines a procedure where MOS is recovered by **bias removal** and **subject inconsistency** weighting. Same procedure is also included in ITU-R P.910 Recommendation Annex-E [7]. The procedure defines the

individual opinion scores of a subject as the combination of subject bias, inconsistency and the true quality of the stimuli and jointly solves these three parameters. Two solvers are proposed for the approach in [1]. Due to minimal differences between the solvers, we only consider the Alternating Projection (AP) solver in this work. P913-12.6 can be seen as the next step of the P913-12.4 by additionally considering subject inconsistency during MOS recovery. Note that the model does not provide any estimate for content ambiguity.

- **MLE:** Li *et al.* [10] proposed a a MOS recovery approach by jointly estimating bias and inconsistency of subject and content ambiguity with Maximum Likelihood Estimation (MLE) and belief propagation. In addition to bias and inconsistency of subject as P913-12.6, MLE also provide content ambiguity. However, it is acknowledged by the authors that the MLE solver has the issue of lacking of uniqueness in its solution in certain cases, *e.g.,* it cannot find solutions for HD-VJND dataset (see Sec.2).

To address the limitations of previous work, we propose an alternative method that relies on Z-score to estimate subject bias, inconsistency, and content ambiguity. We also present a simple yet efficient MOS and POS recovery scheme. Our proposed model is more robust to different use-cases and datasets as it does not require a solver, which can sometimes result in convergence issues. The contributions of this work are:

- A simple yet robust statistical model for estimating subject bias, inconsistency, and content ambiguity from subjective opinion scores.
- A MOS and POS recovery method based on the estimated subject bias and inconsistency.
- Performance comparison between the proposed model and the state-of-the-art, validated by estimating CIs and quantifying the impact of p_{th} POS recovery on the accuracy of SUR prediction models.

2. QOE DATASETS

The performance of ZREC and other existing models from the literature in terms of MOS and POS recovery, as well as estimation of subject bias, inconsistency and content ambiguity, have been assessed on two datasets with distinct characteristics.

HD-VJND: is a Video-Wise Just Noticeable Differences (VW-JND) dataset for HD videos [11]. There are 180 source content (SRC) evaluated by 20 naive subjects with correct visual acuity. Each SRC has been compressed with HEVC with different Constant Rate Factor (CRF) and presented to each subject via Robust Binary Search [11] to find the JND of each subject. Therefore the proxy of JND is represented by CRF value. Satisfied-User-Ratio (SUR) curve is the complementary cumulative distribution function of the individual JNDs of a viewer group [12]. $q\%$SUR is the CRF value that corresponds to a SUR value on the SUR curve equals to threshold $q\%$. 75% is the most commonly used threshold [13, 14, 15, 16, 17]. In this work, the individual JND annotations for each subject are considered as the opinion scores. Additionally, the $q\%$SUR is equivalent to the p_{th} percentile ($p = 1-q$, see Eq.(11) in Sec.3) of opinion score. The opinion scores were used for MOS/CI validation and parameter estimation experiments as well as to measure the impact of POS recovery on the accuracy of SUR prediction models.

Netflix Public: Netflix Public Dataset [18] is a publicly available video quality dataset with 79 Processed Video Sequences (PVS) where each evaluated by 26 subjects. We used the opinion scores for MOS/CI validation and parameter estimation experiments.

3. PROPOSED MODEL

Let $o_{i,j}$ be the opinion score annotated by subject i for stimulus j. For a subjective dataset that consist of m stimulus and have been evaluated by n subjects, the original annotation can be represented by a matrix $\mathbf{O} \in \mathbb{R}^{n\times m}$. For every stimulus, we first compute the mean and standard deviation of the opinion score annotated by each subject:

$$\mathbf{m}(j) = \left(\frac{1}{n}\sum_{i=1}^{n} o_{i,j}\right), \text{where } j = 1, 2, \dots m \tag{1}$$

$$\mathbf{s}(j) = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(o_{i,j} - \mathbf{m}(j)\right)^2}\right), \text{where } j = 1, 2, \dots m \tag{2}$$

where $\mathbf{m}, \mathbf{s} \in \mathbb{R}^{1\times m}$.

Afterwards, we acquire the Z-score matrix $\mathbf{Z}$ from the raw opinion score matrix $\mathbf{O}$ as:

$$\mathbf{Z} = \frac{\mathbf{O} - \mathbf{Im}}{\mathbf{Is}} \tag{3}$$

where $\mathbf{I} = [1, 1, \dots 1]^T$ and $\mathbf{I} \in \mathbb{R}^{n\times 1}$.

Each element $z_{i,j}$ in matrix $\mathbf{Z}$ represents the number of standard deviations by which the opinion score $o_{i,j}$ is away from the m_j. The following analyses are mainly based on the Z-score matrix $\mathbf{Z}$.

3.1. Subject bias and inconsistency

Let $\mathbf{B} \in \mathbb{R}^{1\times n}$ and $\mathbf{C} \in \mathbb{R}^{1\times n}$ the vector of bias and inconsistency of n subjects respectively. Bias and inconsistency for subject i is calculated with the mean and standard deviation of the Z-score for subject i across all stimulus:

$$\mathbf{B}(i) = \left(\frac{1}{m}\sum_{j=1}^{m} z_{i,j}\right), \text{where } i = 1, 2, \dots n \tag{4}$$

$$\mathbf{C}(i) = \left(\sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(z_{i,j} - \mathbf{B}(i)\right)^2}\right), \text{where } i = 1, 2, \dots n \tag{5}$$

The key distinction between the estimation of subject bias in ZREC and P913.12-4 is that ZREC describes the subject bias in the standard deviation range of each stimulus, while P913.12-4 describes it in the opinion score range. By modeling subject bias in the standard deviation range of individual stimuli, ZREC takes stimulus ambiguity into account.

3.2. Content ambiguity

It is important to clarify the difference between stimuli ambiguity and content ambiguity. In QoE datasets, multiple stimuli (*i.e.,* PVS) can be generated from a unique source content (*i.e.,* SRC). We define the stimulus ambiguity for j as the standard deviation of subjects' opinion score (Eq.(2)). Consequently, content ambiguity is defined as the mean ambiguity of all stimuli that belong to a particular content l:

$$\mathbf{A}(l) = \left(\frac{1}{h} \sum_{j \in \mathbf{g}} \mathbf{s}(j) \right), \text{where } l = 1, 2, \dots t \tag{6}$$

h is the number of stimulus for content l, $\mathbf{g}$ the list of stimulus index of content l, t is the total number of contents in the entire datasets, $\mathbf{A} \in \mathbb{R}^{1 \times t}$.

3.3. Mean opinion score recovery

We first remove the bias of each subject for each stimuli from the original annotation $\mathbf{O}$. The unbiased opinion score matrix $\mathbf{U} \in \mathbb{R}^{n \times m}$ is calculated with:

$$\mathbf{U} = \mathbf{O} - \mathbf{B^T s} \tag{7}$$

$u_{i,j}$ is the element of $\mathbf{U}$, which is the opinion score of subject i for stimuli j after the removal of the bias of subject i. In Eq.(7), we multiply the bias of each observers $\mathbf{B^T} \in \mathbb{R}^{n \times 1}$ with the standard deviation $\mathbf{s} \in \mathbb{R}^{1 \times m}$ of different subjects' opinion score for every stimuli in Eq.(2) in order to re-scale the Z-score to the original opinion score range.

To calculate the recovered MOS of stimuli j, denoted $\mathbf{R}(j)$, we employ a weighting scheme that takes into account the inconsistency of the opinion scores provided by different subjects. Specifically, instead of simply averaging the unbiased scores u_j across all subjects, we use a weighted average of u_j, where the weight assigned to each score is inversely proportional to the subject's inconsistency. This means that subjects with higher inconsistency are given less weight, and their opinion scores have less influence on the final MOS calculation.

$$\mathbf{R}(j) = \left(\frac{\sum_{i=1}^{n} \mathbf{C}(i)^{-2} u_{i,j}}{\sum_{i=1}^{n} \mathbf{C}(i)^{-2}} \right), \text{where } j = 1, 2, \dots m \tag{8}$$

Similar with the recovered MOS, weighted standard deviation is calculated as:

$$\boldsymbol{\sigma}_w(j) = \left(\sqrt{\frac{n}{n-1} \times \frac{\sum_{i=1}^{n} \mathbf{C}(i)^{-2} (u_{i,j} - \mathbf{R}(j))^2}{\sum_{i=1}^{n} \mathbf{C}(i)^{-2}}} \right) \tag{9}$$

Where $j = 1, 2, ...m$. The factor of $n/(n-1)$ is intended to account for the number of degrees of freedom, thus giving us an unbiased estimation of the standard deviation of the population[19]. The 95% CI is thus computed with:

$$\mathbf{CI}(j) = \mathbf{R}(j) \pm 1.96 \frac{\sigma_w(j)}{\sqrt{n}} \tag{10}$$

3.4. P_{th} percentile opinion score recovery

Algorithm 1 Calculate Q_p, weighted p_{th} percentile opinion scores

Require: unbiased subject opinions matrix, $U_{n,m}$
Require: subject inconsistencies, C_n
Require: percentile to be calculated, p
  number of subjects = n, number of stimuli = m
  total weight of the population, w = sum(C_n^{-2})
  percentile weight, $w_p = w \times p/100$
  initialize Q_p, a zero vector (with size=m) to store p_{th} percentile opinion score for each stimuli
  for each stimuli j in m **do**
    $U_n \leftarrow$ get subject opinions for stimuli j from $U_{n,m}$
    $U_{n-sorted} \leftarrow$ sort U_n in ascending order
    $w_{n-sorted} \leftarrow$ sort C_n^{-2} with same indices as $U_{n-sorted}$
    initialize current weight, $w_c = 0$
    initialize current subject index, $i = 0$
    while $w_c < w_p$ **do**
      $Q_p(j) \leftarrow$ get current subject (i) opinion from $U_{n-sorted}$ and set it as the p_{th} percentile score
      $w \leftarrow$ get current subject (i) weight from $w_{n-sorted}$
      $w_c \leftarrow w_c + w$
      $i \leftarrow i + 1$
    end while
  end for
  return Q_p

Some subjective/objective studies are not interested in mean of opinion scores but the percentile of opinion scores. For JND and SUR studies [11, 13, 15, 16, 17, 20], 75%SUR is commonly used to train and evaluate objective metric. It can be easily proved that for a given stimuli j:

$$q\%\text{SUR}(o_i) = (1-q) \text{ - } th \text{ percentile}(o_i), \tag{11}$$

75%SUR is in fact 25_{th} percentile. Therefore, we provide a weighted percentile approach where subject bias and inconsistency is taken into account. Algorithm 1 depicts the process to calculate weighted percentiles Q_p of an unbiased opinion score matrix $U_{n,m}$ of n subjects and m stimuli for a given percentile p in range [0, 100].

Table 2. Analysis of the CI of the recovered MOS for four different methods on the Netflix Public and the HD-VJND datasets. Avg CI represents the length of the CI for each method on each dataset with all subjects included in MOS recovery. CI% represents the percentage of recovered MOS values that fall within the confidence interval range based on 1000 bootstrapping iterations. In each iteration, only half of the subjects in each dataset are used to recover the MOS.

	NETFLIX		HD-VJND	
	Avg CI	CI%	Avg CI	CI%
BT500	0.5153	0.5645	1.2612	0.9285
P913-12.4	0.4986	0.9102	1.1254	0.8671
P913-12.6	0.4420	0.8885	1.0217	0.8805
ZREC	0.4172	0.8783	0.9813	0.8554

Table 3. Pearson linear correlation coefficient (PLCC) between the estimated parameters of subject inconsistency, subject bias, and content ambiguity across various models.

	Subject Inconsistency	Subject Bias	Content Ambiguity
Model Pair	NETFLIX		
MLE - ZREC	0.9282	0.9952	0.9663
MLE - P913.12-6	0.9669	0.9964	-
P913.12-6 - ZREC	0.9372	0.9965	-
P913.12-4 - MLE	-	0.9992	-
P913.12-4 - P913.12-6	-	0.9999	-
P913.12-4 - ZREC	-	0.9965	-
Model Pair	HD-VJND		
P913.12-6 - ZREC	0.9603	0.9994	-
P913.12-4 - P913.12-6	-	0.9999	-
P913.12-4 - ZREC	-	0.9994	-

4. EXPERIMENT RESULTS

4.1. MOS recovery and confidence intervals

Table 2 depicts the average 95% CI of the recovered MOS on the Netflix Public and HD-VJND datasets with all subjects. To evaluate the comparative reliability of the confidence intervals generated by each method, we performed a bootstrapping analysis comprising 1000 iterations. In each iteration, we randomly selected half of the subjects and recovered the MOS with each model. Results shows that ZREC and P913.12-6 exhibit the lowest average CI values while maintaining a relatively high CI% level. Despite having a higher CI% value, P913.12-4 displays a significantly larger average CI compared to ZREC and P913.12-6.

4.2. Estimated parameters

In this section, we analyze the correlation between the subject bias, inconsistency and content ambiguity across the tested models. As summarized in the Table 1, BT500 does not estimate any of the parameters and thus excluded from the correlation analysis. Moreover, P913.12.4 cannot estimate subject inconsistency and content ambiguity while P913.12.6 cannot estimate content ambiguity. In addition, MLE fails to converge to a solution for HD-VJND dataset.

Table 3 depicts the PLCC values between the indicated model pairs in each row. The results indicate that the tested models are well correlated in terms of subject bias. On the other hand, estimated subject inconsistencies show slight differences between models. Finally, MLE and ZREC shows relatively high correlations in terms of content ambiguity. Despite the lower correlations for subject inconsistencies, ZREC estimations are in line with the standards. It is impossible to know which model estimations are closer to the ground truth, however the analysis showcases the relative reliability of the approach.

Table 4. The mean and variance of absolute errors on 75%SUR prediction with SUR prediction model [11] on HD-VJND dataset without any recovery and with ZREC and P913.12-6 POS recovery.

$\lvert\Delta 75\%\text{SUR}\rvert$	Without Recovery[11]	P913-12.6 POS Recovery	ZREC POS Recovery
Mean Error	0.7489	0.7175	**0.6883**
Error Variance	0.9224	0.7198	**0.6989**

4.3. Impact of percentile opinion score recovery on the accuracy of SUR prediction models

Previous work [9] has shown that training objective quality models on cleaned data can improve the prediction performance. In this work, we compared the performance of the 75%SUR prediction model [11] trained on 75%SUR from original datasets without recovery and 75%SUR (25_{th} percentile) recovered by ZREC and P913-12.6 respectively. Because P913-12.6 only provide MOS recovery but not percentile recovery, we use Algorithm 1 with the subject bias and inconsistency of P913-12.6 as input. The mean and variance of absolute error of 75%SUR for different training data are shown in Table 4. It can be observed that the 75%SUR prediction model trained both on ZREC and P913-12.6 improved the prediction, in which ZREC get a smaller prediction error than P913-12.6.

5. CONCLUSION

We introduced ZREC to estimate subject bias, inconsistency and content ambiguity, all of which are fundamental for QoE studies. Using these parameters, ZREC can recover the MOS and the POS whichever is more suitable for the QoE use-case in question. Our findings indicate that ZREC can produce slightly tighter CIs for MOS recovery on two datasets compared to the current state of the art models, albeit with a minor reduction in accuracy. A tighter CI allows to reduce the required number of subjects in the subjective study without sacrificing from the accuracy and resolving power. Furthermore, the results of our experiments on the SUR prediction use-case demonstrate that ZREC can improve the performance of objective quality metrics by providing a more reliable ground truth with 25_{th} POS recovery.

6. REFERENCES

[1] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," in *Human Vision and Electronic Imaging 2020, Burlingame, CA, USA, 26-30 January 2020*, Ingenta, 2020.

[2] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, (New York, NY, USA), p. 3339–3347, Association for Computing Machinery, 2020.

[3] J. Li, S. Ling, J. Wang, Z. Li, and P. L. Callet, "Gpm: A generic probabilistic model to recover annotator's behavior and ground truth labeling," 2020.

[4] A. Ak, A. Goswami, W. Hauser, P. Le Callet, and F. Dufaux, "Rv-tmo: Large-scale dataset for subjective quality assessment of tone mapped images," *IEEE Transactions on Multimedia*, pp. 1–12, 2022.

[5] A. Ak, M. Abid, M. Perreira Da Silva, and P. Le Callet, "On spammer detection in crowdsourcing pairwise comparison tasks: Case study on two multimedia qoe assessment scenarios," pp. 1–6, 07 2021.

[6] ITU-R, "Methodology for the subjective assessment of the quality of television pictures." ITU-R Recommendation BT.500-14, 2019.

[7] ITU-R, "Subjective video quality assessment methods for multimedia applications." ITU-R Recommendation Recommendation P.910, 2022.

[8] ITU-R, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment." ITU-R Recommendation Recommendation P.913, 2021.

[9] A.-F. Perrin, C. Dormeval, Y. Wang, N. Birkbeck, B. Adsumilli, and P. L. Callet, "When is the cleaning of subjective data relevant to train ugc video quality metrics?," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 1466–1470, 2022.

[10] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data compression conference (DCC)*, pp. 52–61, IEEE, 2017.

[11] J. Zhu, A.-F. Perrin, and P. Le Callet, "Subjective test methodology optimization and prediction framework for just noticeable difference and satisfied user ratio for compressed hd video," in *2022 Picture Coding Symposium (PCS)*, pp. 313–317, IEEE, 2022.

[12] J. Zhu, P. Le Callet, A.-F. Perrin, S. Sethuraman, and K. Rahul, "On the benefit of parameter-driven approaches for the modeling and the prediction of satisfied user ratio for compressed video," in *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 4213–4217, IEEE, 2022.

[13] X. Zhang, C. Yang, H. Wang, W. Xu, and C.-C. J. Kuo, "Satisfied-user-ratio modeling for compressed video," *IEEE Transactions on Image Processing*, vol. 29, pp. 3777–3789, 2020.

[14] H. Wang, I. Katsavounidis, Q. Huang, X. Zhou, and C.-C. J. Kuo, "Prediction of satisfied user ratio for compressed video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6747–6751, IEEE, 2018.

[15] Y. Zhang, H. Liu, Y. Yang, X. Fan, S. Kwong, and C. J. Kuo, "Deep learning based just noticeable difference and perceptual quality prediction models for compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1197–1212, 2021.

[16] H. Lin, V. Hosu, C. Fan, Y. Zhang, Y. Mu, R. Hamzaoui, and D. Saupe, "Sur-featnet: Predicting the satisfied user ratio curve for image compression with deep feature learning," *Quality and User Experience*, vol. 5, pp. 1–23, 2020.

[17] C. Fan, H. Lin, V. Hosu, Y. Zhang, Q. Jiang, R. Hamzaoui, and D. Saupe, "Sur-net: Predicting the satisfied user ratio curve for image compression with deep learning," in *2019 eleventh international conference on quality of multimedia experience (QoMEX)*, pp. 1–6, IEEE, 2019.

[18] "Netflix public dataset.https://github.com/netflix/vmaf/blob/master/resource/doc/datasets.md," Online; accessed 20-Feb-2023.

[19] P. R. Bevington, D. K. Robinson, J. M. Blair, A. J. Mallinckrodt, and S. McKay, "Data reduction and error analysis for the physical sciences," *Computers in Physics*, vol. 7, no. 4, pp. 415–416, 1993.

[20] H. Wang, X. Zhang, C. Yang, and C.-C. J. Kuo, "Analysis and prediction of jnd-based video quality model," in *2018 picture coding symposium (PCS)*, pp. 278–282, IEEE, 2018.