Subjective Media Quality Recovery From Noisy Raw Opinion Scores: A Non-Parametric Perspective

Andrés Altieri, Lohic Fotio Tiotsop, *Member, IEEE*, and Giuseppe Valenzise, *Senior Member, IEEE*

***Abstract*—This paper focuses on the challenge of accurately estimating the subjective quality of multimedia content from noisy opinion scores gathered from end-users. State-of-the-art methods rely on parametric statistical models to capture the subject's scoring behavior and recover quality estimates. However, these approaches have limitations, as they often require restrictive assumptions to achieve numerical stability during parameter estimation, leading to a lack of robustness when the modeling hypotheses do not fit the data. To overcome these limitations, we propose a paradigm shift towards non-parametric statistical methods. Specifically, we introduce a threefold contribution: i) in contrast to the prevailing approach in subjective quality recovery assuming a parametric score distribution, we propose a non parametric approach that guarantees greater accuracy by measuring reliability per subject and per stimulus, overcoming the limits of existing approaches that measure only per subject reliability; ii) we propose ESQR, a non-parametric algorithm for subjective quality recovery, demonstrating experimentally that it has higher robustness to noise compared to numerous state-of-the-art algorithms, thanks to the weaker assumptions made on data compared to parametric approaches; iii) the proposed approach is theoretically grounded, i.e., we define a non-parametric statistic and prove mathematically that it provides a measure of score reliability.**

***Index Terms*—Multimedia quality assessment, non-parametric method, opinion scores reliability, subjective quality recovery.**

I. Introduction

SUBJECTIVE tests are evaluations conducted by human observers to assess the quality and usability of multimedia applications, such as audio, video, or graphical content. These tests involve participants providing their opinions, preferences, or judgments based on their personal experiences and perceptions. Data gathered in subjective tests serve as a crucial resource for advancing the design of multimedia applications. These data are expected to provide insight into end-users' overall satisfaction, thereby guiding the development and optimization of multimedia technologies to better meet user needs and expectations.

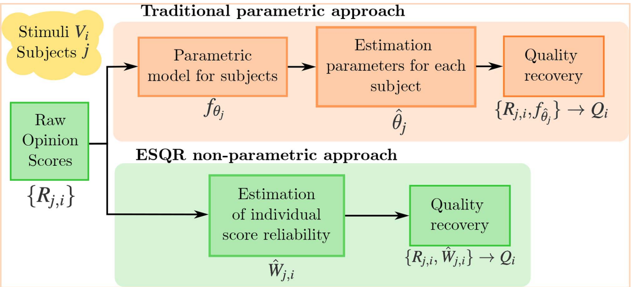

Traditional parametric approach
Stimuli V_i Subjects j
Parametric model for subjects
f_{θ_j}
Estimation parameters for each subject
$\hat{\theta}_j$
Quality recovery
$\{R_{j,i}, f_{\hat{\theta}_j}\} \to Q_i$
Raw Opinion Scores
$\{R_{j,i}\}$
ESQR non-parametric approach
Estimation of individual score reliability
$\hat{W}_{j,i}$
Quality recovery
$\{R_{j,i}, \hat{W}_{j,i}\} \to Q_i$

Fig. 1. Parametric approaches assume that the observed opinion scores come from a probabilistic model f_θ, that depends on a finite set of parameters θ. The parameters θ are then estimated, and the subjective quality Q is recovered from the assumed probabilistic model using the estimated parameters. The proposed non-parametric approach (ESQR) instead measures how reliable is each single opinion score without resorting to a predefined model. The recovered quality Q is then computed as a weighted sum of the opinion scores in which the weight of each opinion score is determined by its reliability.

However, a challenge arises from the inherent noise present in raw data collected during subjective tests. This noise stems from various factors, including subjects' fatigue and distraction, unforeseen software issues, and other uncontrolled elements influencing the emotional state of the subjects. Consequently, the raw data collected in a subjective test may not precisely reflect the end-users' satisfaction with the multimedia application under evaluation. This challenge has motivated the need for approaches to recover the subjective quality of multimedia content from noisy raw data collected from subjective tests.

Several approaches have been developed to recover an accurate estimate of the subjectively perceived media quality from noisy raw ratings collected in subjective tests [1], [2], [3], [4]. The conventional approach computes the mean of the opinion scores (MOS) gathered for a given stimulus as an estimate of the subjective quality of that stimulus. Despite its popularity, the MOS is known to be particularly sensitive to outlier opinion scores [5]. As a consequence, more sophisticated approaches that exploit parametric statistical models have recently been proposed for the subjective media quality recovery problem [3], [6], [7], [8].

The diagram in the top part of Fig. 1 illustrates the main steps of parametric approaches (e.g., [1], [3], [6], [8], [9]). They essentially implement three steps: i) a parametric statistical model explaining the scoring behavior of each subject is assumed; ii) the model's parameters are estimated; iii) finally the desired

Manuscript received 29 December 2023; revised 18 March 2024; accepted 7 April 2024. Date of publication 17 April 2024; date of current version 26 August 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mai Xu. *(Corresponding author: Lohic Fotio Tiotsop.)*

Andrés Altieri is with the CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, Université Paris-Saclay, 91190 Gif-sur-Yvette, France, and also with CSC-CONICET, Buenos Aires C1425FQD, Argentina (e-mail: andres.altieri@centralesupelec.fr).

Lohic Fotio Tiotsop is with the Department of Control and Computer Engineering, Corso Duca degli Abruzzi 24, Politecnico di Torino, 10129 Torino, Italy (e-mail: lohic.fotiotiotsop@polito.it).

Giuseppe Valenzise is with the CNRS, CentraleSupelec, Laboratoire des Signaux et Systèmes, Université Paris-Saclay, 91190 Gif-sur-Yvette, France (e-mail: giuseppe.valenzise@centralesupelec.fr).

The code to run ESQR and reproduce the results in this paper is made freely available at: http://media.polito.it/ESQR.

Digital Object Identifier 10.1109/TMM.2024.3390113

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

subjective quality is derived. The main advantage of parametric approaches is that they explain the scoring behavior of subjects. However, the use of parametric approaches entails several limitations: i) Lack of robustness, as the model's assumptions that allow interpretability and numerical stability in practice do not hold in several application scenarios [6], [8]. ii) High risk of underfitting since despite the huge number of factors that influence the subjects scoring behavior, very few parameters are considered to preserve the model's numerical stability. For instance, authors of parametric approaches (e.g., [1], [3], [6]) very often use a single parameter to capture the reliability of a subject, disregarding the fact that the latter varies with the characteristics of each stimulus that the subject is asked to evaluate. iii) High computational complexity due to the parameter estimation process that involves solving a complex optimization problem [3], [7].

Despite the acknowledged limitations of parametric statistical approaches for multimedia quality recovery, there has been a notable absence of research exploring a shift to non-parametric statistical methods which are less sensitive to the aforementioned modeling issues of parametric methods. This paper fills this gap by investigating such a paradigm shift for the first time. More precisely, this paper proposes the first *non-parametric* approach to study users' scoring behavior in subjective quality assessment. The general scheme of the proposed method is illustrated in the bottom part of Fig. 1. The proposed non-parametric approach considers two main steps: first, it measures the reliability of each single opinion score without resorting to any assumed probabilistic scoring model. Then, the recovered quality is computed as a weighted sum of the opinion scores, in which the weight of each opinion score is determined by its reliability.

The proposed approach to measure the reliability of an opinion score finds its theoretical explanation in information theory. In fact, it is inversely proportional to a measure of how surprising that opinion score is for the quality of the stimulus under evaluation. The proposed method addresses the main shortcomings of parametric approaches since it: i) does not consider any scoring model involving simple yet restrictive assumptions on the subject's behavior; ii) does not require to solve an optimization problem for parameters estimation, thus a significant computational burden is avoided; iii) does not suffer underfitting or overfitting issues since it is a non-parametric approach.

The contributions of this paper can be summarized as follows:

1) We introduce the first non-parametric approach for measuring subject reliability in subjective experiments. While existing parametric methods assess reliability solely on a per-subject basis, our approach extends to measuring reliability both per subject and stimulus. This accounts for the variability in scoring behavior observed across different stimuli rated by subjects, thereby providing a more comprehensive understanding of reliability in subjective quality assessment.
2) We propose ESQR, a novel non-parametric algorithm for subjective media quality recovery. By adopting a non-parametric perspective, ESQR offers enhanced robustness compared to current state-of-the-art approaches which all rely on parametric models. As already mentioned, existing parametric models make assumptions about subject scoring behavior to ensure numerical stability during parameter estimation. ESQR's departure from such assumptions marks a significant step forward in developing more accurate and adaptable subjective quality recovery algorithms.
3) Lastly, the proposed approach is theoretically grounded, i.e., we define a non-parametric statistic and prove mathematically that it provides a measure of score reliability. This theoretical underpinning represents a departure from the prevailing trend in the field, where quality recovery approaches rely on scoring models assumed a priori.

Computational experiments demonstrate that ESQR offers a subjective quality estimate with reduced uncertainty compared to current techniques. Additionally, ESQR exhibits lower sensitivity to noise when compared to five state-of-the-art quality recovery approaches. Furthermore, the findings suggest as expected that the performance of ESQR is relatively robust across various application scenarios. This resilience could be attributed to the fact that, unlike parametric approaches, our method circumvents assumptions about subjects' scoring behavior that may be invalid in certain application contexts.

The rest of this paper is organized as it follows: Section II reviews and analyzes the prior art. Our non-parametric approach to analyze subject's behavior in subjective tests is described in Section III, while the proposed quality recovery algorithm is discussed in Section IV. Computational experiments and the related results are presented in Section V, and the conclusions are drawn in Section VI.

II. Related Work

In recent years, the research on subjective quality recovery has gained increasing importance, specially since the COVID-19 pandemic democratized the collection of opinion scores in non-highly controlled environments, i.e., crowdsourcing subjective tests [10], [11], [12]. Various scholars have delved into the exploration of factors contributing to the presence of noise on raw individual ratings derived from subjective experiments. These factors encompass the influence of experimental context [13], the impact of subject fatigue [14], and instances where subjects misunderstand the task, potentially leading to inverted ratings [15].

In efforts to mitigate the influence of noise sources on raw opinion scores, numerous methodologies have been scrutinized for the subjective assessment of media content quality [16]. These methodologies include single stimulus-based approaches [17], where subjects exclusively evaluate the processed signal; pair comparisons [18], involving subjects comparing the quality of stimulus A to that of stimulus B and indicating the superior one; and double stimulus-based approaches [17], wherein subjects rate the quality of processed content relative to reference content after viewing the latter.

Empirical observations suggest that pair comparison-based subjective experiments are prone to yield more accurate results compared to single stimulus-based experiments [19]. However, the comprehensive acquisition of comparison matrices demands considerable time, posing challenges to conducting pair comparison-based experiments with numerous stimuli, as feasible in single stimulus-based approaches. Furthermore, while double stimulus-based methods yield quality scores with narrower confidence intervals, they necessitate twice the time of single stimulus approaches. Consequently, data collection approaches aimed at enhancing the precision of raw ratings entail constraints on the maximum number of stimuli that can be evaluated.

Beyond the pragmatic limitations associated with approaches likely to ensure greater accuracy of opinion scores, there exist noise sources beyond the researcher's control during subjective testing. For instance, a subject's rating for a specific video sequence may significantly reflect their personal content preferences, such as the scene's likability or dislikability [20].

Several approaches to recover the subjective quality from raw opinion scores have been proposed in the literature. Examples of approaches to deal with noise in pair comparison-based subjective tests can be found, e.g., in [21], [22], [23], [24]. Pairwise preferences and rating scores can also be fused into a common quality scale, thus reducing the noise of individual experiments [25], [26]. In this paper, we focus on the scenario where a rating quality scale is used to elicit and collect opinions.

The mean of the opinion scores (MOS) gathered for a given stimulus has long been considered as a good estimate of the subjective quality of that stimulus. Unfortunately, the MOS is very sensitive to outlier ratings since it attributes the same importance to reliable and unreliable subjects. To address this limitation, several relevant ITU recommendations have appeared. The ITU-R BT.500 [5] recommends to identify outlier subjects and exclude their opinion scores from the dataset before computing the MOS. The ITU-T P.910 [27] introduces the so-called Absolute Category Rating (ACR) scale and recommends the use of the quality recovery algorithm proposed in ITU-R BT.500 [5] together with Confidence Intervals (CIs) to deal with noisy raw opinion scores collected using the ACR scale. Finally, the ITU-T P.913 [28] suggests to first perform a bias removal step on the data and then potentially use the algorithm proposed in ITU-R BT.500 [5].

All these ITU recommendations support the use of an algorithm that performs subject exclusion. However, removing all the opinion scores of a subject from the dataset probably causes an unnecessary loss of relevant information. In fact, as underlined in [3], [6], considering that the excluded subjects gave inaccurate opinion scores for all the stimuli is overly conservative. Therefore, researchers have recently proposed advanced statistical approaches to circumvent subject exclusion.

These recent methods mainly assume that each opinion score of each subject follows a probability distribution that can be characterized by a finite number of parameters. These parameters are then estimated using statistical frameworks such as the Maximum Likelihood Estimation (MLE) [29] or the Expectation Maximization Algorithm (EMA) [30]. Specifically, most of these recent parametric approaches assume that the scoring behavior of a subject can be modeled with two parameters, i.e., the subject's bias and inconsistency [1], [6], [8], [9], [31]. The bias is defined as a systematic tendency of a subject to provide smaller (negative bias) or larger (positive bias) opinion scores than the actual subjective quality of the stimulus that is being rated. The inconsistency instead is a measure of the inability of a subject to provide consistent opinion scores when rating the same stimulus more than once.

When relying on bias and inconsistency, the authors usually assume that each raw opinion score of each subject is a realization of a Gaussian random variable. In [1], [9], the mean of such a Gaussian random variable was modeled as the sum of two parameters, i.e., the subjective quality to be recovered and the subject's bias. The variance was also expressed as the sum of two other parameters: the subject's inconsistency and the stimulus's ambiguity. The authors of [6] argued that the Gaussian model proposed in [1] is still valuable if one gets rid of the stimulus's ambiguity, thus the variance of the Gaussian random variable modeling each opinion score of each subject was assumed to be equal only to the subject's inconsistency. The authors then proposed an iterative algorithm, called alternating projection, to recover the subjective quality based on their proposed scoring model. The alternating projection algorithm was implemented in the Netflix SUREAL software [32]. The authors of [8] improved the Gaussian model of [6] by integrating into it the so-called Standard deviation of Opinion Scores (SOS) hypothesis, proposed by the authors of [33], in order to account for the fact that subjects are more accurate when rating very low or very high quality. Finally, the authors of [31] suggested to subtract the MOS from the raw opinion scores and divide the result by the SOS to obtain the so-called Z-scores. They then argued that the bias, the inconsistency and the subjective quality can be estimated more reliably and efficie ntly from the Z-scores rather than the original opinion scores. This yielded a subjective quality recovery approach called "ZREC".

Another interesting parametric approach is the one proposed in [3]. The authors assumed that each rating of each subject is a realization of a mixture of two probability distributions. The first probability distribution models accurate opinion scores given by the subject, while the second accounts for the cases in which the subject provides inaccurate opinion scores. The accuracy of the subject is measured with one parameter, i.e., the probability that the subject would score the quality according to the distribution modeling accurate opinion scores. An EMA is proposed by the authors to estimate the model's parameters and thus the subjective quality.

More recently, the authors of [7] introduced the so-called Regularized MLE (RMLE) approach to recover subjective media quality. Basically, they proposed a regularization term to be added to the likelihood function before solving the optimization problem, the solution of which provides an estimate of the parameters that characterize the subjective quality to be recovered. The proposed regularization term is meant to penalize opinion scores that are potentially noisy.

Although the superiority over the MOS of the parametric approaches discussed so far has already been proved in specific applications, the assumptions made by these approaches on the scoring behavior of subjects raises some questions on their general applicability. In fact, as it will be observed in Section V, the performance of some of these approaches with respect to the performance of the MOS can vary significantly from one application scenario to another. In addition, from a theoretical point of view, the optimization problem guiding the parameter estimation process in parametric models is usually computationally challenging and might not even have a unique optimal solution as pointed out in [31]. This introduces additional challenges that parametric models have to overcome.

This paper adopts a different perspective than the recent parametric approaches described in the previous paragraphs. We avoid making assumptions about any particular parametric probabilistic model able to explain all the opinion scores of a subject. Instead, a non-parametric approach to measure the reliability of each opinion score of each subject and recover the subjective quality of each stimulus is proposed. By refraining from imposing stringent assumptions on the subjects' scoring behavior, we aim at proposing an approach whose accuracy is preserved over a broad spectrum of practical situations.

The setting of this article shares some common elements with the problem of crowd-source learning or learning from noisy labels [34], [35], which has drawn considerable attention in recent years, specially in the context of classification problems [36], [37], [38], [39], [40], [41], [42], [43]. A common aspect that we see between the problem of performing crowd-source learning and the subjective media quality recovery problem that we consider in this paper is that in both cases, the input is a set of noisy human-generated labels. However, there are some key differences in terms of goals and how the performance of an approach is assessed in both fields. When performing crowd-source learning, the goal is to propose an algorithm which effectively uses the crowdsource labels to mitigate the effect of the noise during the training process. In contrast, in subjective quality recovery, the ground truth subjective quality corresponds to a latent and unobservable random variable. The immediate goal is to propose an algorithm that derives, from noisy crowdsource labels, a robust point estimates of specific moments of this latent random variable or its probability distribution in the most general case.

In addition, to measure the performance of approaches aiming at crowd-source learning, there is typically a ground-truth test set with correct labels, on which the trained algorithm can be tested to report a prediction accuracy. In contrast, in subjective quality recovery, there is no ground-truth test set, the performance of subjective quality recovery algorithms has to be determined through different approaches that have been validated in the literature, such as robustness to perturbations, size of confidence intervals and also through the intrinsic theoretical properties of the proposed estimates. The unavailability of test sets is due to the intrinsic subjectiveness of the media quality assessment task. A label cannot be unequivocally identified as right or wrong, and the variability within the labels represents different preferences and expectations within a population. Approaches such as majority voting [44] are not suitable for this problem since they disregard such a variability. The representativeness of an opinion score has to be evaluated in terms of its value within the population and the observed behavior of the subject.

III. A Non-Parametric Measure of Reliability

This section introduces and motivates our approach to measure the reliability of each opinion score of each subject. The proposed measure will be used later in Section IV to derive a new algorithm to recover the subjective media quality.

As discussed in the previous section, parametric approaches to recover the subjective media quality from raw opinion scores typically consider some parameters that measure the reliability of each subject during the subjective test. For instance, in [6], [31], the inverse of the square of the subject's inconsistency was used as measure of reliability. The authors of [3] estimated the probability of each subject to provide accurate scores, and then used this probability as an indicator of the reliability of the subject. Clearly, the effectiveness of these measures of reliability strongly depends on the assumptions of the underlying parametric model. In this paper, we argue that the reliability of each opinion score can be effectively measured without resorting to any parametric scoring model.

A. *Notation and Hypotheses*

Let us introduce the following notation:

- $\mathcal{I}$, a set of rated stimuli.
- $\mathcal{J}$, a set of subjects.
- $\mathcal{I}_j \subset \mathcal{I}$ the subset of the stimuli rated by the subject $j \in \mathcal{J}$.
- $\mathcal{J}_i \subset \mathcal{J}$, the subset of subjects that rated the stimulus $i \in \mathcal{I}$ using a discrete scale in the range $\{1, \ldots, K\}$.

For a given stimulus $i \in \mathcal{I}$, there might be different opinion scores that accurately characterize its quality. The ground-truth quality can be modeled as a discrete random variable, characterized by its probability mass function (pmf). Specifically, we introduce the following notation:

- V_i the discrete random variable that describes the latent ground-truth quality of a stimulus $i \in \mathcal{I}$ in the range $\{1, \ldots, K\}$. In the absence of any noise, subjects sample this random variable, producing opinion scores.
- p_{V_i} denotes the pmf of V_i.

The problem of quality estimation has been formulated in some cases as that of recovering p_{V_i} (refer to, e.g., [3], [45]). In practice, most of the quality assessment literature focuses on predicting point estimates of the quality pmf, such as the mean opinion score. We also follow the latter approach in this paper, i.e., we are interested in estimating $q_i = \mathbf{E}[V_i]$.

In real-world scenarios, the latent random variable describing quality is observed through a set of stochastic, *noisy* observations, i.e., opinion scores. This model effectively captures the intrinsic subjectivity of scores, arising, e.g., from different expectations in terms of quality. We will therefore denote by:

- $R_{j,i}$, the discrete random variable modeling the score of the subject j for the stimulus i on a quality scale in the range $\{1, .., K\}$.
- $p_{R_{j,i}}$ denotes the pmf of $R_{j,i}$.

Our goal is to propose a robust estimator of q_i by introducing an approach to measure the reliability of each observed opinion score $\{R_{j,i}\}_{j \in \mathcal{I}_j, i \in \mathcal{I}}$. We will denote as Q_i our proposed estimator of q_i.

B. Measuring the Reliability of an Opinion Score

We propose the following definition of reliability:

Definition 1: The reliability $W_{j,i}$ of the opinion score $R_{j,i}$ given by the subject j when assessing the quality of the stimulus i is the following ratio:

$$W_{j,i} = -\frac{1}{\log\left(p_{V_i}(R_{j,i})\right)}. \tag{1}$$

An estimate of the distribution p_{V_i} is required to compute the reliability measures $W_{j,i}$. We will explore how to perform such an estimation in Section IV. For the moment, in order to motivate why the formula in (1) provides a suitable measure of reliability for each opinion score gathered in a subjective test, let us assume that the distribution p_{V_i} is known for each stimulus $i \in \mathcal{I}$.

Note that $W_{j,i}$ is a measure of reliability if and only $W_{j,i}^{-1}$ is a measure of unreliability. To motivate the suitability of $W_{j,i}$ as a measure of reliability, let us introduce the following statistic for each subject:

$$S_j(\mathcal{I}_j) = \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} W_{j,i}^{-1}. \tag{2}$$

We will prove that the statistic $S_j(\mathcal{I}_j)$ can be used to measure the average unreliability of the subject j. In fact, the following proposition holds:

Proposition 1: For each subject j, if there is a constant c such that $\text{var}[W_{j,i}^{-1}] < c \, \forall i \in \mathcal{I}_j$, then, as $|\mathcal{I}_j| \to \infty$, $S_j(\mathcal{I}_j)$ converges in the mean square sense to:

$$\bar{h}_j(\mathcal{I}_j) = \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} H(p_{R_{j,i}}) + \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} D_{\text{KL}}(p_{R_{j,i}}||p_{V_i}), \tag{3}$$

where $H(p_{R_{j,i}})$ is the entropy of the distribution $p_{R_{j,i}}$ and $D_{\text{KL}}(p_{R_{j,i}}||p_{V_i})$ denotes the Kullback-Leibler (KL) divergence between $p_{R_{j,i}}$ and p_{V_i}.

Proof: We should prove that:

$$\lim_{|\mathcal{I}_j|\to\infty} \mathbb{E}\left[\left(S_j(\mathcal{I}_j) - \bar{h}_j(\mathcal{I}_j)\right)^2\right] = 0. \tag{4}$$

To do this we show that $S_j(\mathcal{I}_j)$ is an unbiased estimator of $\bar{h}_j(\mathcal{I}_j)$ and that its variance goes to zero as $|\mathcal{I}_j| \to \infty$.

For the bias we have:

$$\begin{aligned}
\mathbb{E}_{R_{j,i}}[S_j(\mathcal{I}_j)] &= -\frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j}\mathbb{E}_{R_{j,i}}[\log(p_{V_i}(R_{j,i})] \\
&= -\frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j}\mathbb{E}_{R_{j,i}}\left[\log\left(\frac{p_{V_i}(R_{j,i})}{p_{R_{j,i}}(R_{j,i})}\right)\right] \\
&\quad - \mathbb{E}_{R_{j,i}}\left[\log\left(p_{R_{j,i}}(R_{j,i})\right)\right] \qquad (5)\\
&= \bar{h}_j(\mathcal{I}_j), \qquad (6)
\end{aligned}$$

where from (5) to (6), we use the definitions of entropy and KL divergence. Also, we consider that if there are any terms in the expectation for which $p_{R_{j,i}}(r) = 0$, then those terms are zero through the continuity definition that $0 \log 0 = 0$ [46], so in reality we are only dividing by $p_{R_{j,i}}$ for the positive terms that impact the expectation.

Finally, to compute the variance of $S_j(\mathcal{I}_j)$ we make use of the fact that the opinion scores can be considered independent and the assumption that the variance of $W_{j,i}^{-1}$ is finite:

$$\text{var}\left[S_j(\mathcal{I}_j)\right] = \frac{1}{|\mathcal{I}_j|^2}\sum_{i\in\mathcal{I}_j}\text{var}\left[W_{j,i}^{-1}\right] \le \frac{c}{|\mathcal{I}_j|} \xrightarrow[|\mathcal{I}_j|\to\infty]{} 0, \tag{7}$$

which completes the proof. □

Proposition 1 basically states that the statistic $S_j(\mathcal{I}_j)$ converges to the overall unreliability of the subject j, since for a large value of $|\mathcal{I}_j|$, the following approximation can be used:

$$S_j(\mathcal{I}_j) \approx \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} H(p_{R_{j,i}}) + \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} D_{\text{KL}}(p_{R_{j,i}}||p_{V_i}). \tag{8}$$

Remembering that the subject's inconsistency is a measure of the inability to repeat exactly the same opinion score when rating several times the same stimulus, it is not difficult to observe that the more the subject j is inconsistent when rating the stimulus i, the larger is the entropy $H(p_{R_{j,i}})$ of the random variable $R_{j,i}$. Hence, the quantity $\frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} H(p_{R_{j,i}})$ in (8) captures the average *inconsistency* of the subject j. On the other hand, the following $\frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} D_{\text{KL}}(p_{R_{j,i}}||p_{V_i})$ indicates on average how far the opinion scores of the subject j are expected to be from the accurate opinion scores. Hence, $\frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} D_{\text{KL}}(p_{R_{j,i}}||p_{V_i})$ measures the average *inaccuracy* of the subject j.

In the light of the above interpretation of Proposition 1 it turns out that asymptotically the value of the statistic $S_j(\mathcal{I}_j) = \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} W_{j,i}^{-1}$ measures the average *unreliability* of the subject j. Each single term, i.e., $W_{j,i}^{-1}$, of this statistic can therefore be considered as the measure of the unreliability of the opinion score given by the subject j to the quality of the stimulus i. This motivates our definition of $W_{j,i}$ as a measure of reliability.

It is worth noting that in information theory, the logarithm of the probability of an event is called the self-information of the event. It can be interpreted as a measure of the likelihood of the event, and thus the level of surprise that the observation of the event entails. Therefore, $-\log(p_{V_i}(R_{j,i}))$ can be interpreted as the measure of how surprising (thus potentially unreliable) is the opinion score $R_{j,i}$ if it is supposed to come from p_{V_i}. This is another way to motivate why the $W_{j,i}$ defined in (1) measures the reliability of the opinion score $R_{j,i}$.

Notice that Proposition 1, which is the main theoretical foundation of the proposed reliability measure, is based on the notion of entropy of a probability distribution (the KL divergence is the relative entropy between two probability distributions). For this reason, the algorithm proposed in the next section, grounded in the proposed reliability measure, shall be denoted as **Entropy-based Subjective Quality Recovery (ESQR)**.

IV. Recovering the Subjective Quality: The ESQR Algorithm

We now introduce the proposed ESQR algorithm to recover the subjective quality. The key idea of the method is to weight each opinion score in the computation of the subjective media quality using the reliability measure described in Section III.

To measure the reliability of each opinion score using the formula in (1), the distribution of accurate opinion scores p_{V_i} of each stimulus i is required. In practice, as already mentioned, this distribution is unknown, so an estimate of it is needed. Therefore, the ESQR algorithm implements two main steps:

1) For each stimulus i, an estimate $\hat{p}_{V_i}$ of the distribution of accurate opinion scores is obtained from the observed sample of opinion scores $\{R_{j,i}\}_{i \in \mathcal{I}_j, j \in \mathcal{J}}$.
2) Then, $\hat{p}_{V_i}$ is used in (1) to get an estimate $\hat{W}_{j,i}$ of the reliability of each opinion score. $\hat{W}_{j,i}$ is used to define our estimator of the ground-truth subjective quality q_i of each stimulus i as it follows:

$$Q_i = \frac{\sum_{j \in \mathcal{J}_i} \hat{W}_{j,i}.R_{j,i}}{\sum_{k \in \mathcal{J}_i} \hat{W}_{k,i}}. \tag{9}$$

Note that from the definition of Q_i above, the contribution of the subject $j \in \mathcal{J}_i$ to the determination of the ground truth quality of the stimulus $i \in \mathcal{I}$ is weighted by their normalized reliability i.e.,

$$\omega_{ij} = \frac{\hat{W}_{j,i}}{\sum_{k \in \mathcal{J}_i} \hat{W}_{k,i}}. \tag{10}$$

A. Estimating p_{V_i}

Unlike previous approaches, we do not make assumptions on the shape of the distributions p_{V_i}, $i \in \mathcal{I}$. Therefore, we use a non-parametric approach to estimate the distribution of accurate opinion scores.

The simplest non-parametric estimation of p_{V_i} is based on computing the histogram of opinion scores. In fact, $p_{R_{j,i}}$ can be estimated by the histogram of the opinion scores that the subject j gave for the quality of the stimulus i. For instance, if a single opinion score is collected per stimulus, then $p_{R_{j,i}}$ is a probability mass function that attributes a one to the observed opinion score and 0 to all the other opinion scores on the quality scale. The observed distribution of opinion scores for each stimulus i can then be estimated as:

$$\bar{p}_{V_i}(r) = \frac{1}{|\mathcal{J}_i|} \sum_{j \in \mathcal{J}_i} p_{R_{j,i}}. \tag{11}$$

In practice, this estimate $\bar{p}_{V_i}$ can be influenced by the potential presence of noise in the raw opinion scores collected from the subjects. To obtain a more accurate estimation that closely aligns with the true distribution p_{V_i} we suggest a different approach for weighting different subjects' contributions, as opposed to the uniform weighting $(1/|\mathcal{J}_i|)$ used in (11). Specifically, our proposition is based on the observation that the more a subject's opinion scores correlate with those of other subjects, the more trustworthy that subject is. Consequently, to enhance the precision of our estimate of p_{V_i} we propose applying a weighting coefficient, denoted as ϵ_j to each histogram $p_{R_{j,i}}$ which depends on the overall correlation between subject j and the other subjects.

Algorithm 1: Entropy Based Subjective Quality Recovery (ESQR).

```
Data: $R_{j,i},\ i \in \mathcal{I}_j;\ j \in \mathcal{J}$ // stimuli $i$, subjects $j$
    $C_{jk} \leftarrow \mathrm{SROCC}(R_{j,.}, R_{k,.})\ j,k \in \mathcal{J}$ // pairwise subject scores correlation
    $\hat{C}_j \leftarrow \mathrm{FZT}^{-1}\left(\frac{\sum_{k \in \mathcal{J}} \mathrm{FZT}(C_{jk})}{|\mathcal{J}|}\right)\ j \in \mathcal{J}$ // overall subject-to-subject correlation
    $\epsilon_{ij} \leftarrow \frac{|\hat{C}_j|}{\sum_{k \in \mathcal{J}_i} |\hat{C}_k|}\ i \in \mathcal{I};\ j \in \mathcal{J}_i$ // importance of the ratings of subject $j$ in the $P_{V_i}$ estimation
    $\hat{p}_{V_i} \leftarrow \sum_{j \in \mathcal{J}_i} \epsilon_{ij} p_{R_{j,i}}\ i \in \mathcal{I}$ // estimate the distribution $P_{V_i}$
    $\hat{W}_{j,i} \leftarrow \frac{1}{-\log(\hat{p}_{V_i}(R_{j,i}))}\ i \in \mathcal{I}_j;\ j \in \mathcal{J}$ // estimate each opinion's score reliability
    $Q_i \leftarrow \frac{\sum_{j \in \mathcal{J}_i} \hat{W}_{j,i} R_{j,i}}{\sum_{k \in \mathcal{J}_i} \hat{W}_{k,i}}\ i \in \mathcal{I}$ // estimate the quality
Result: $Q_i,\ i \in \mathcal{I}$
```

We employ C_{jk}, the Spearman Rank Order Correlation Coefficient (SROCC) between the opinion scores of the subject j and those of the subject k, as a non-parametric measure of correlation. To compute the average correlation between subject j's opinion scores and those of all the other subjects, we utilize the Fisher Z-Transformation (FZT), as suggested in [47]. For each subject j, the FZT is applied to the SROCC values $C_{jk}\quad k = 1, 2, \ldots, j-1, \ldots, j+1, \ldots, |\mathcal{J}_i|$. The average of the obtained values is computed. The inverse of the FZT, here denoted by FZT^{-1}, is then applied to the obtained average to obtain the overall correlation $\hat{C}_j$ between the opinion scores of the subject j and those of the other subjects. Finally, the importance ϵ_{ij} of the histogram $p_{R_{j,i}}$ in the estimation of the distribution p_{V_i} is expressed as:

$$\epsilon_{ij} = \frac{|\hat{C}_j|}{\sum_{k \in \mathcal{J}_i} |\hat{C}_k|} \quad i \in \mathcal{I}, \quad j \in 1, 2, \ldots, |\mathcal{J}_i|. \tag{12}$$

Therefore, the final estimate $\hat{p}_{V_i}$ of the distribution p_{V_i}, used to recover the subjective quality of each stimulus i as defined in (9), is obtained from the following formula:

$$\hat{p}_{V_i} = \sum_{j \in \mathcal{J}_i} \epsilon_{ij}.p_{R_{j,i}} \quad i \in \mathcal{I}. \tag{13}$$

It is worth noticing that other more sophisticated estimations of $\hat{p}_{V_i}$ might be possible. However, we argue that even a simple and possibly noisy approximation of p_{V_i}, such as the one proposed above, is a good starting point for the second step of the proposed ESQR algorithm, where atypical subjects are further penalized using the reliability measure introduced in Section III. The experiments in Section V support this claim with empirical evidence.

The proposed ESQR algorithm is summarized in Algorithm 1. The input of Algorithm 1 is the set of the observed opinion scores $R_{j,i}\ i \in \mathcal{I}_j;\ j \in \mathcal{J}$. The output is the recovered subjective quality of each stimulus i. The notation $R_{j,.}$ is used to indicate all the opinion scores of the subject j.

B. Confidence Interval of the Recovered Quality

Using a similar formula as in [31] to compute the standard deviation of a weighted sum of opinion scores, for each stimulus i, an unbiased estimator of the standard deviation of Q_i can be computed as follows:

$$\sigma_{Q_i} = \sqrt{\frac{|\mathcal{J}_i|}{|\mathcal{J}_i| - 1} \frac{\sum_{j \in \mathcal{J}_i} \hat{W}_{j,i} (R_{j,i} - Q_i)^2}{\sum_{k \in \mathcal{J}_i} \hat{W}_{k,i}}}. \tag{14}$$

From the standard deviation in (14), the 95% confidence interval (CI) of the recovered quality stimulus i can be computed as:

$$\mathrm{CI}_{Q_i} = Q_i \pm 1.96 \frac{\sigma_{Q_i}}{\sqrt{|\mathcal{J}_i|}}. \tag{15}$$

Equation (15) assumes that the estimator Q_i is normally distributed. In Appendix A, we provide theoretical conditions for this to happen, and we show through simulation that these conditions are indeed reasonable in a practical scenario. Also, notice that this normality assumption only applies to the *estimator* of the ground-truth quality, and not the individual opinion scores, which can follow any arbitrary distribution in our framework.

V. Numerical Experiments

A. Experimental Settings

The evaluation of subjective media quality recovery methods is challenging, since there is no observable "true" quality of stimuli to be used as ground truth. In related work, e.g., [1], [6], [7], [8], [31], the effectiveness of quality recovery approaches is assessed in terms of: i) robustness to the insertion of synthetic noise in the quality scores; ii) and uncertainty on the recovered subjective media quality. We will therefore use similar experiments in this paper to evaluate the effectiveness of the proposed ESQR algorithm.

We compare the proposed ESQR algorithm with five state-of-the-art quality recovery approaches, i.e.: the MOS, the algorithm recommended in the ITU-R BT.500 [5], very recent algorithms such as ZREC [31], RMLE [7], The Generalized Distribution Score (GSD)-based approach presented in [48], the SOS Hypothesis-aware Subjective Quality Recovery (SHaSQR) algorithm proposed in [8] and finally the Netflix SUREAL software that implements the so-called "alternating projection" algorithm [6]. The latter has been recommended by the ITU in 2021 as the most comprehensive method for subjective quality recovery (as per Section 12.6 of ITU-R P.913 [28]).

The computational experiments were conducted using the data gathered in six different subjective tests. The related datasets are named: VQEG-HD1 [49], VQEG-HD3 [49], VQEG-HD5 [49], Netflix Public [1], KoNViD-1k [50] and the MoviesLens-1M [51]. While the last two datasets, i.e., the KoNViD-1k [50] and the MoviesLens-1M [51] were obtained from crowdsourcing subjective tests, the others are the results of highly controlled lab experiments. Notice that for crowdsourcing experiments, the matrix of opinion scores is typically sparse, as stimuli are evaluated only by a subset of subjects. Thus, we present results for crowdsourcing datasets in a separate section below.

TABLE I
Uncertainty of Quality Estimates: Comparison of the Size of CIs Estimated by the Different Quality Recovery Approaches

Methods	AVG CI SIZE			
	NETF PUB	**VQ-HD1**	**VQ-HD3**	**VQ-HD5**
MOS	0.509 (——)	0.493 (——)	0.565 (——)	0.575 (——)
BT500	0.515 (+1.18%)	0.613 (+24.34%)	0.586 (+3.72%)	0.575 (+0.00%)
ZREC	0.417 (-18.07%)	0.437 (-11.36%)	0.458 (-18.94%)	0.475 (-17.39%)
SUREAL	0.445 (-12.57%)	0.459 (-6.90%)	0.481 (-14.87%)	0.489 (-14.96%)
SHaSQR	0.399 (-21.61%)	0.418 (-15.21%)	0.456 (-19.29%)	0.466 (-18.96%)
RMLE	0.453 (-11.00%)	0.417 (-15.42%)	0.472 (-16.46%)	0.483 (-16.00%)
ESQR	**0.355 (-30.26%)**	**0.361 (-26.77%)**	**0.436 (-22.83%)**	**0.439 (-23.65%)**

Percentages indicate relative size of the CIs with respect to MOS CIs.
"The best performance for each testing condition is highlighted in bold".

For the three VQEG experiments, there were 24 participants and each of them rated around 168 stimuli, yielding for each of the three tests, a total of 24 × 168 opinion scores to be analyzed. The Netflix Public dataset is a relatively small-scale dataset, which includes the opinion scores of 26 subjects on the perceptual quality of 70 processed video sequences and nine source content. The KoNViD-1 k subjective test involves 624 participants, who have scored 1200 short video sequences. For the MovieLens-1 M, 6040 subjects have expressed their opinion score on 3952 movies. For all the six datasets considered for our experiments, the authors made use of five-point quality scales when gathering the opinion scores from the subjects. Hence, the opinion scores in each dataset range from 1 to 5.

B. Uncertainty of Quality Estimates

A typical approach to measure the uncertainty of the subjective quality recovered by a given method consists in computing the size of confidence intervals [6], [8], [31]. The larger the CI, the higher the uncertainty on the recovered subjective quality.

Table I shows the comparison between the average size of the CIs of the recovered subjective quality by each method on the four datasets resulting from tests performed in controlled environments. The percentages reported between parenthesis indicate by how much the application of each method reduced on average the size of the CIs that can be computed from the raw opinion scores, i.e. the MOS's CIs (computed for each stimulus i as MOS $\pm 1.96 \times$ SOS/$\sqrt{|\mathcal{J}_i|}$, where SOS stands for Standard deviation of Opinion Scores). For instance, as it can be seen in Table I, the average of the sizes of the MOS's CIs on the Netflix public dataset is 0.509, while the average of the sizes of the CIs of the recovered qualities by the proposed ESQR algorithm is 0.355. Hence, by using the proposed ESQR algorithm instead of the MOS on the Netflix public dataset, on average, the size of the CIs of the recovered subjective qualities is reduced by 30%, i.e., $100 \times (1 - 0.355/0.509)$.

Looking at the results in Table I, it can be noticed that on all datasets, the proposed ESQR algorithm always recovered subjective qualities characterized by smaller CIs than those of all the other approaches on average. Hence, in practice, the proposed algorithm is expected to provide estimates of the subjective quality that are prone to lower uncertainty.

It is interesting to notice that more recent approaches such as SHaSQR, ZREC, RMLE and the Netflix SUREAL software offered better performances than the MOS and the algorithm proposed in the ITU-R BT.500. The performances of the Netflix SUREAL software, RMLE, ZREC and SHaSQR were however outperformed by that of the proposed ESQR algorithm. In fact, the application of the ESQR algorithm has yielded in the worst case a reduction of the size of CIs by more than 22% in all cases, while all the other approaches never did better than 21%. The CIs resulting from the output of the algorithm proposed in the ITU-R BT.500 are indeed larger than the MOS's CIs on average. This is actually not a peculiarity of this work as the same observation was made. We also note that the GSD approach was not considered in this experiment since we did not find in the related paper a formula to compute CIs for the recovered qualities. in [6].

C. CIs Prediction Accuracy

When comparing confidence interval sizes, a natural question arises: does a smaller confidence interval actually imply reduced uncertainty, or is it merely a result of underestimating the true uncertainty linked to the quality estimation? In real datasets, there are no ground-truth CIs against which estimated CIs can be benchmarked. Therefore, we must resort to simulations to verify the accuracy the CI estimates of ESQR and competing methods.

We simulated the opinion scores of 25 subjects for 100 stimuli. For each stimulus i, we assumed that reliable opinion scores on its quality follow a normal distribution with a mean of q_i and a standard deviation of σ_i. Consequently, q_i represents the ground truth quality for stimulus i. The ground truth CI of the quality of stimulus i is then:

$$\mathrm{CI}_i = q_i \pm 1.96 \frac{\sigma i}{\sqrt{M_s}}, \tag{16}$$

where $M_s = 25$ is the number of simulated opinion scores for each stimulus. The ground truth quality values q_i were derived by uniformly sampling 100 numbers within the range of [1.5, 4.5]. To simulate the fact that subjects exhibit lower inconsistency at the quality scale's extremes, as observed in real subjective experiments [33], we employed the SOS hypothesis [33]. More precisely, we set $\sigma_i = 0.2 \times (-q_i^2 + 6q_i - 5)$, ensuring that the standard deviation of the distribution of reliable opinion scores diminishes at the quality scale's extremes.

We will denote $N(q_i, \sigma_i)$ as the distribution of reliable opinion scores for the stimulus i. In our simulation, each stimulus is assessed by 25 subjects. We followed the scoring model proposed in [3], where each subject could provide a reliable opinion score with a probability of $1 - \eta$ and an unreliable one with probability η. We divided the 25 subjects into two clusters, i.e., a group of 20 accurate subjects, and a group of 5 inaccurate ones. For the accurate subjects, we set $\eta = 0.01$ (1%), meaning that 99% of their opinion scores were sampled from the distribution $N(q_i, \sigma_i)$ of reliable opinion scores and rounded to the closest integer from 1 and 5, while the remaining 1% were randomly selected between 1 and 5. The 5 inaccurate subjects had η randomly chosen between 0.6 and 1, meaning that at least 60% of their opinion scores were randomly selected between 1 and 5, and the rest were drawn from $N(q_i, \sigma_i)$. We conducted this simulation with 30 different seeds, resulting in 30 distinct simulated datasets.

TABLE II
CI PREDICTION ACCURACY

Method	MOS	BT500	ZREC	SUREAL	SHaSQR	RMLE	ESQR
Δ^m	0.13	0.06	**0.05**	**0.05**	0.08	0.08	**0.05**
ρ^m	1.47	1.26	1.24	1.24	1.21	1.25	**0.98**

"The best performance for each testing condition is highlighted in bold".

We applied all quality recovery methods to each of the simulated datasets. Let $\hat{\mathrm{CI}}_{id}^m$ represent the CI estimated by method m for stimulus i in simulated dataset d. To evaluate the accuracy of method m in estimating the ground truth CIs, we compared $\hat{\mathrm{CI}}_{id}^m$ to CI_i using two main indices:

$$\Delta^m = \frac{1}{30 \times 100} \sum_{d=1}^{30} \sum_{i=1}^{100} |ct(\hat{\mathrm{CI}}_{id}^m) - ct(\mathrm{CI}_i)| \tag{17}$$

$$\rho^m = \frac{1}{30 \times 100} \sum_{d=1}^{30} \sum_{i=1}^{100} sz(\hat{\mathrm{CI}}_{id}^m)/sz(\mathrm{CI}_i) \tag{18}$$

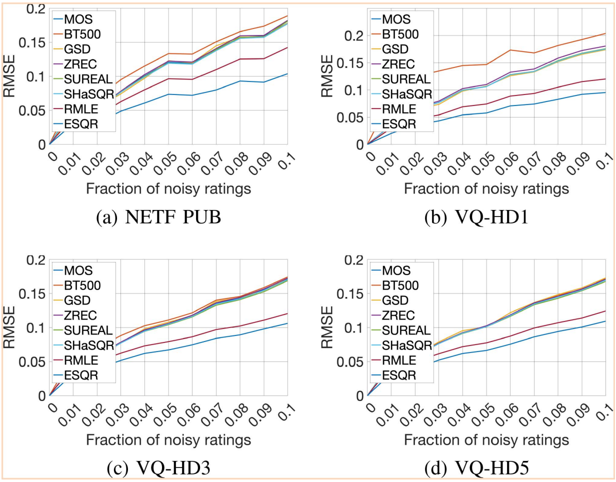
where $ct()$ and $sz()$ stand for center and size of the CI respectively. Δ^m is, therefore, the average distance between the center of the estimated CI of method m and the center of the ground truth CI. Meanwhile, ρ^m is the average ratio between the size of the CI estimated by method m and the size of the ground truth CI. Clearly, the closer Δ^m is to zero, the better; and the closer ρ^m is to 1, the better.

Table II summarizes the results. Regarding Δ^m, the best methods are ESQR, ZREC, and Netflix Sureal software, with the center of the estimated CI differing from that of the ground truth CI by around 0.05. The MOS exhibited the lowest performance ($\Delta^m = 0.13$), followed by RMLE and SHaSQR ($\Delta^m = 0.08$). When it comes to predicting CI sizes, ESQR outperformed all other approaches, with a related value of $\rho^m = 0.98$ significantly closer to 1 than that of all the other methods for which $\rho^m > 1.21$. Thus, ESQR slightly underestimated (by 2% of the actual size) the sizes of the ground truth CIs on average, while all the other methods overestimated them significantly (by more than 21% of the actual size).

The results in Table II suggest that the proposed ESQR algorithm can better predict the ground truth CIs and thus better quantify the actual uncertainty characterizing the quality of a stimulus compared to the other quality recovery approaches. We believe this stems from the fact that ESQR makes no restrictive assumptions about the subjects' scoring behavior.

D. Robustness to Synthetic Noise

Following the same protocol of [1], [6], [7], [8], [31], all the quality recovery methods are first used to recover the subjective media quality on each dataset. Then, some synthetic noise is added to each dataset. After adding the noise, the subjective quality is estimated again, this time using the noisy dataset. Finally, the recovered quality on the noisy dataset is compared in terms of RMSE to the one obtained before adding noise to



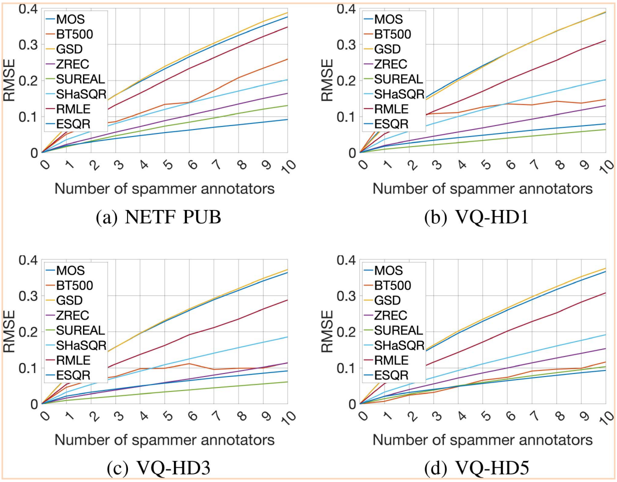
(a) NETF PUB

(b) VQ-HD1

(c) VQ-HD3

(d) VQ-HD5

Fig. 2. Robustness to noise insertion. RMSE between the quality recovered on the original dataset and under different noisy conditions. The noise was added by replacing a given fraction of the opinion scores (see the x-axis) of each subject with integer numbers sampled at random between 1 and 5. The simulation was run with 30 different seeds and the curve for each quality recovery method reports the average RMSE from the 30 trials.



(a) NETF PUB

(b) VQ-HD1

(c) VQ-HD3

(d) VQ-HD5

Fig. 3. Robustness to spammer annotators. RMSE between the quality recovered on the original dataset and under noisy conditions. The noise was generated by adding simulated subjects (see the x-axis) that score the quality of each stimulus with an integer number sampled at random between 1 and 5. The simulation was run with 30 different seeds and RMSE of the 30 trials for each quality recovery method is shown.

the dataset. This allows us to determine which method is more robust to noise.

The noise is added to the datasets using two different approaches that simulate different applications:

1) *Noise insertion:* a small fraction of the opinion scores of each subject is replaced by an integer sampled at random in the interval $[1, 5]$;
2) *Spammer annotators*: some simulated subjects scoring the quality at random are added to the dataset.

In practice, our first approach to adding noise simulates, for instance, the type of noise that would be generated by the subjects' fatigue or unexpected subjects' distraction. In fact, it is reasonable to assume that, due to fatigue or distraction, each subject might inaccurately score the quality of a very small fraction of stimuli. The second noise model simulates situations such as the unexpected crash of the software used to collect opinion scores, causing a mismatch of the opinion scores of certain subjects (see for instance the Netflix subjective test described in [1]). In that case, the subjects whose opinion scores have been mismatched correspond to subjects rating at random. Another application is the case in which a subject accepts to participate in the subjective test but just provides ratings at random in order to complete the test as quickly as possible. These subjects are referred to in the literature as spammer annotators [15].

Fig. 2 reports the results obtained using the first noise model. The fraction of replaced opinion scores is reported on the x-axis. The y-axis reports, for each quality recovery method, the RMSE error between the quality recovered on the original dataset and the one obtained after replacing the fraction of opinion scores on the x-axis with random integers. For instance, looking at the VQEG-HD1 dataset in Fig. 2, the RMSE between the values of the recovered qualities by the proposed ESQR algorithm on the original dataset and the values computed after replacing 4% of the opinion scores of each subject in the dataset with random integers is 0.06.

In this first case, as it can be seen from Fig. 2, the proposed ESQR algorithm outperformed all the other methods in all testing conditions. In fact, the curve of RMSE values associated to the ESQR algorithm lies below the ones of all the other methods. This result suggests that the proposed ESQR algorithm would guarantee better robustness than the other quality recovery methods to the noise generated for instance by the subjects' fatigue.

One can notice that the Netflix SUREAL software, SHaSQR, GSD and ZREC showed performances very similar to that of the MOS. The RMLE approach instead showed better performance than the MOS, the BT.500 algorithm, ZREC,SHaSQR, GSD and the Netflix SUREAL software.

The results related to our second approach to add the noise are summarized in Fig. 3. The x-axis reports the number of simulated subjects added to the dataset. These simulated subjects rate the stimuli by choosing at random an integer between 1 and 5. In this case, the proposed ESQR algorithm, ZREC and the Netflix SUREAL software showed comparable performances. In particular ESQR delivered the best performance on the Netflix public dataset and provided similar performance to that of the Netflix SUREAL software on the VQEG-HD5 dataset. The Netflix SUREAL software showed a better performance than the proposed ESQR algorithm on the VQEG-HD1 and VQEG-HD3 datasets. The RMLE approach showed a performance that is higher than that of the MOS but significantly lower than those of all the other methods. Finally, it can be observed that the GSD approach is not particularly robust to the presence of spammer annotators. This can be explained by the fact that, unlike SUREAL, ZREC and SHaSQR, the GSD was mainly proposed to better model the distribution of opinion scores stimulus by stimulus, rather than capturing long term inaccuracy that characterizes spammer annotators.

The joint analysis of Figs. 2 and 3 reveals a crucial observation: the proposed ESQR algorithm demonstrates robustness, irrespective of the noise simulation approach. Unlike other quality recovery methods tested, ESQR performance does not exhibit sensitivity to the specific noise simulation employed. For instance, in Fig. 2, sophisticated methods like ZREC, SHaSQR and the Netflix SUREAL software do not outperform the MOS, yet they perform well in the second noise simulation case, as depicted in Fig. 3. Conversely, RMLE excels in the first case in Fig. 2 but falters in the second case in Fig. 3. ESQR consistently maintains high performance across both cases, standing out as the top performer in the first noise simulation case and one of the best methods in the second case.

This stability in ESQR's performance during transitions between scenarios can be attributed to our avoidance of assumptions about subjects' scoring behavior, a characteristic of parametric approaches. Such assumptions often face challenges due to specific application characteristics. For example, the parametric model in the Netflix SUREAL software assumes a subject permanently possesses bias and inconsistency, making it less effective in capturing the scoring behavior of subjects who only occasionally misjudge quality. This likely explains the similar performance of the Netflix SUREAL software to that of the MOS in the first noise simulation case.

E. Crowdsourcing Experiments

This section evaluates the accuracy of ESQR when the matrix of ratings is sparse. This type of matrix is typically obtained from crowdsourcing tests where a very large number of stimuli is employed, but each subject is required to rate only a small subset of them. This yields a stimuli-to-subjects table with numerous empty cells and thus a sparse matrix of ratings.

When the matrix of ratings is sparse, the correlation between the ratings of each pair of subjects cannot always be calculated. This makes it difficult for ESQR to derive a more accurate estimate of the distribution p_{V_i} than the distribution of collected ratings. In this case, our implementation of ESQR considers the distribution $\bar{p}_{V_i}$ (11) of gathered ratings during the test as the estimate of p_{V_i} for each stimulus i in order to compute the reliability of each individual opinion score.

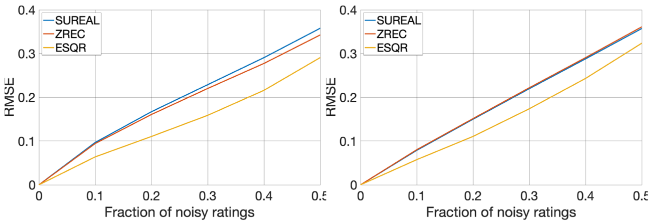
In light of the results discussed in Section V-D, the quality recovery approaches with the most competitive performance with respect to the proposed ESQR algorithm in terms of robustness to noise are the RMLE (see Fig. 2), ZREC and the Netflix SUREAL software (see Fig. 3). The analysis in this section could have therefore been done by considering ZREC, the Netflix SUREAL software and the RMLE approach. Unfortunately the RMLE approach involves an optimization problem whose solution is computationally very demanding on large-scale datasets as the ones considered in this section. This made it impossible to perform the experiments with RMLE on these datasets in a reasonable amount of time. For this reason, we considered only ZREC and the Netflix SUREAL software.

The sizes of the CIs of the recovered subjective qualities by the proposed ESQR algorithm, ZREC and the Netflix SUREAL software are compared in Table III. As in the case of datasets collected in controlled environments, the proposed ESQR algorithm provided a recovered subjective quality prone to lower uncertainty, i.e. smaller CIs. The use of the Netflix SUREAL software and ZREC induced no more than 18% reduction of the size of the raw data CIs, while ESQR achieved 25%. The reduction percentages achieved by the Netflix SUREAL software are slightly higher than the percentages it reached on datasets obtained in controlled environments (see Table I). This is consistent with the fact that greater benefit can be expected from sophisticated quality recovery approaches when used on challenging datasets such as those derived from crowdsourcing subjective tests.

TABLE III
CROWDSOURCING EXPERIMENTS: CI SIZES OF ESQR VS NETFLIX, SUREAL, ZREC. PERCENTAGE REDUCTION WITH RESPECT TO MOS CI

Methods	AVG CI SIZE	
	KoNViD-1k	**MoviesLens-1M**
SUREAL	0.326 (−15.76%)	0.203 (−15.25%)
ZREC	0.318 (−18.04%)	0.205 (−14.58%)
ESQR	**0.289 (−25.51%)**	**0.195 (−18.75%)**

"The best performance for each testing condition is highlighted in bold".

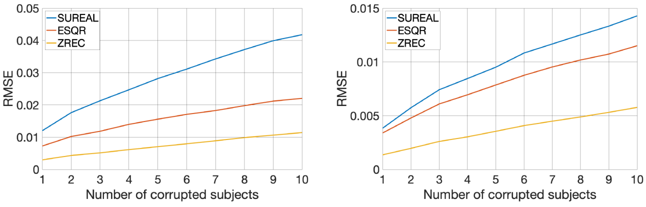


(a) KoNViD-1k (b) MoviesLens-1M

Fig. 4. Crowdsourcing experiments, robustness to noise insertion. Comparing the robustness to synthetic noise of the ESQR algorithm, the Netflix SUREAL software and ZREC on two crowdsourcing datasets. A given fraction of the opinion scores (see the x-axis) in each dataset was replaced with integers sampled at random between 1 and 5. The RMSE between the quality recovered on the original dataset and the one obtained in each noisy situation is reported on the y-axis.

The reduction percentages obtained for the proposed ESQR algorithm in Table III, although being greater than those of the Netflix SUREAL software and ZREC, were in one case smaller than the ones in Table I, which were obtained on datasets collected in controlled environments. This is because the current version of the proposed ESQR algorithm to analyze a sparse matrix of ratings directly uses the distribution of gathered ratings to estimate the reliability of individual opinion scores. We strongly believe that, as for the Netflix SUREAL software, the application of the ESQR algorithm would bring larger benefits on crowdsourcing datasets if an approach to "clean" the distribution of collected opinion scores is employed as in the case of a plain matrix of ratings where pairwise correlations are used. Finding such an approach will thus be one of the main points for a future contribution.

We compared the proposed ESQR algorithm to the Netflix SUREAL software and ZREC in terms of robustness to synthetic noise and spammer annotators added to a sparse matrix of ratings. The results are shown in Figs. 4 and 5. As in Section V-D, the quality recovered on the original dataset was compared in terms of RMSE to the one recovered from a noisy version of



(a) KoNViD-1k (b) MoviesLens-1M

Fig. 5. Crowdsourcing experiments, robustness to spammer annotators. Comparing the robustness of the ESQR algorithm, the Netflix SUREAL software and ZREC to the insertion of spammer annotators on two crowdsourcing datasets. A certain number of spammer annotators (see the x-axis) was added to each dataset. The opinion scores of a spammer annotators are integers sampled at random between 1 and 5. The RMSE between the quality recovered on the original dataset and the one obtained after adding spammer annotators is reported on the y-axis.

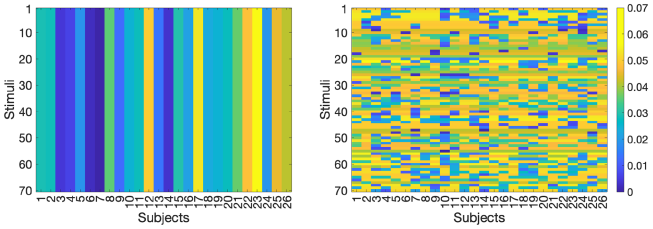


Fig. 6. Reliability of opinion scores. The figure shows the contribution weights $\omega_{ij}^{\text{SUREAL}}$ (left) and $\omega_{ij}^{\text{ESQR}}$ (right) of each subject j to the determination of the ground truth quality of each processed video sequence i in the Netflix public dataset. The ESQR weights can capture the reliability of individual opinion scores.

each dataset. To add noise, a fraction (see the x-axis) of opinion scores was selected at random and replaced with integers randomly sampled between 1 and 5. The ratings of a spammer annotator are simulated by selecting random integers between 1 and 5. As we see in Fig. 4, in the case of noise insertion, the proposed ESQR algorithm showed higher robustness to the added noise. In fact, it always recovered a subjective quality from the noisy dataset with the lowest RMSE with respect to the one obtained on the original version of the dataset. This suggests that, by adding additional noise to a challenging dataset, the proposed ESQR algorithm would offer more robustness to it than the Netflix SUREAL software and ZREC. For what concerns the insertion of spammer annotators (see Fig. 5), ESQR shows better performance than the Netflix SUREAL software, but this performance is outperformed by that of ZREC.

F. Reliability of Opinion Scores

In this section we analyze the effect of the proposed reliability measure (1) as a weight of the contribution of each stimulus to the quality recovery. We compare the proposed reliability weights with the weights defined by Netflix SUREAL and ZREC. Let $\omega_{ij}^{\text{SUREAL}}$, $\omega_{ij}^{\text{ZREC}}$ and $\omega_{ij}^{\text{ESQR}}$ be the weights of the contribution of subject j to the determination of the quality of each stimulus i for SUREAL, ZREC and the proposed ESQR algorithm, respectively. Equation (10) defines $\omega_{ij}^{\text{ESQR}}$, while SUREAL and ZREC define the contribution as:

$$\omega_{ij}^{\text{SUREAL}} = \frac{\left(\sigma_j^{\text{SUREAL}}\right)^{-2}}{\sum_{k\in\mathcal{J}}\left(\sigma_k^{\text{SUREAL}}\right)^{-2}} \tag{19}$$

$$\omega_{ij}^{\text{ZREC}} = \frac{\left(\sigma_j^{\text{ZREC}}\right)^{-2}}{\sum_{k\in\mathcal{J}}\left(\sigma_k^{\text{ZREC}}\right)^{-2}} \tag{20}$$

where σ_j^{SUREAL} and σ_j^{ZREC} are the estimated inconsistency of the subject j by the Netflix SUREAL software and ZREC respectively. Notice that, for SUREAL and ZREC, the weights only depend on the subject and are constant across stimuli evaluated by the same subject.

Equations (19) and (20) reveal that ZREC and SUREAL weigh opinions similarly in determining ground truth quality. Consequently, we exclusively report results comparing ESQR to SUREAL in the following section, as similar conclusions arise with ZREC.

In Fig. 6, we report $\omega_{ij}^{\text{SUREAL}}$ and $\omega_{ij}^{\text{ESQR}}$ computed on the Netflix public dataset. The ability of ESQR weights to modulate the importance of opinion scores *per stimulus* and not only per subject as SUREAL gives higher flexibility and precision in quality estimation. For instance, from the left heatmap in Fig. 6 one can notice that, according to SUREAL, all the opinion scores of the subject #7 are considered unreliable. The heatmap of ESQR contradicts that by identifying stimuli for which the opinion scores of subject #7 are still accurate enough to contribute to estimating ground truth quality. In fact, for the following stimuli: #3, #43, #56 and #58 the ESQR contribution weights of subject #7 are rather high. A quick look at the dataset revealed that for these stimuli, subject #7 gave the opinion score chosen by the majority of subjects. Hence, contrarily to SUREAL, ESQR rightly attributed high importance to these opinion scores since they can be considered as accurate. Similarly, looking at the left heatmap in Fig. 6, it can be observed that SUREAL attributes very high importance to all the opinion scores of subject #23. The ESQR heatmap however points out some stimuli for which the opinion scores of that subject are less reliable. For instance, subject #23 is the only one that scored the quality of stimulus #44 as being “bad” while all the other subjects found it at least “fair”.

Another interesting example is the situation of stimulus #19 for which all the subjects gave the same opinion scores. Despite all the subject agreed on one opinion score, SUREAL attributed different importance to the subjects when recovering the quality of that stimulus. This is not the case for ESQR that attributed the same contribution weight to all subjects as it can be seen from the right heatmap in Fig. 6.

All these examples provide insights of why the proposed ESQR approach achieves in general better robustness and lower uncertainty than SUREAL and ZREC in the numerical experiments presented so far in this paper.

To further assess the effectiveness of the proposed nonparametric measure of reliability, we conducted a simulation study. The objective of the simulation was to demonstrate that as the probability of an opinion score being anomalous increases, the proposed reliability measure progressively identifies such opinion score as less reliable.

In our simulation, consistent with our previous methodology outlined in Section V-C, we assume that reliable opinion scores

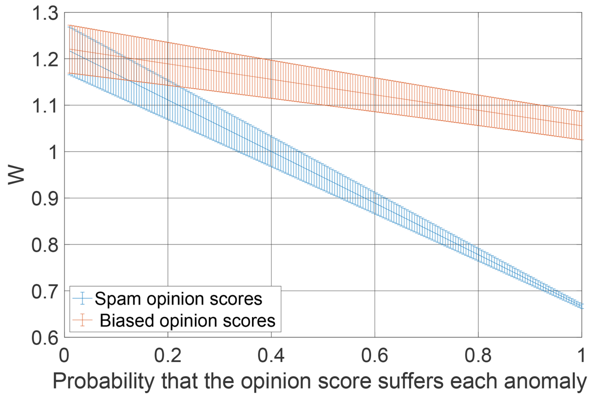


Fig. 7. Average reliability of an opinion score and the related 95% confidence interval as function of the probability that the opinion score is anomalous. Two different anomalies are studied (biased and spam opinion scores). In both cases the proposed measure of reliability decreases as the probability of the opinion scores being affected by an anomaly increases.

regarding the quality of stimulus i follow a normal distribution with a mean of q_i and a standard deviation computed from the SOS hypothesis [33], i.e., $\sigma(q_i) = 0.2 \times (-q_i^2 - 6q_i + 5)$. Thus, the distribution p_{V_i} of reliable opinion scores is denoted as $N(q_i, \sigma_i)$ as in Section V-C. Following the scoring model introduced in [3], we generated the opinion score $R_{j,i}$ that a generic subject j would express on the quality of stimulus i by considering that this opinion score might be reliable with probability $1 - \eta_{ij}$ or anomalous with probability η_{ij}. Consequently, the opinion score $R_{j,i}$ is a realization of a random variable whose distribution is given by the following mixture: $(1 - \eta_{ij}) \times N(q_i, \sigma(q_i)) + \eta_{ij} \times p_Z$, where p_Z represents the probability distribution modeling the type of anomaly affecting the opinion score. In our simulation, we utilized two distinct distributions for Z:

1) The uniform distribution on a discrete scale ranging from 1 to 5. In this case, with probability η_{ij}, the opinion score $R_{j,i}$ is obtained by randomly selecting an integer from 1 to 5. As already mentioned, this type of opinion score is commonly referred to as a "spam" in the literature.
2) A normal distribution with mean $q_i + b$ and standard deviation $\sigma(q_i + b)$. Here, the opinion score is affected by a bias set to b. In our simulation, we employed $b = \pm 0.5$.

Subsequently, we examined the reliability of the opinion score $R_{j,i}$ as a function of its probability η_{ij} of being anomalous.

The results of the simulation are presented in Fig. 7. For each value of η_{ij}, we simulated 100 different opinion scores $R_{j,i}$ by randomly selecting 100 different values of q_i within the interval [1, 5]. This approach enabled us to obtain 100 distinct values of reliability W_{ij} corresponding to a specific η_{ij}. With this sample of values, we computed not only the average reliability corresponding to a given η_{ij} but also the associated 95% CI.

As depicted in Fig. 7, as the probability η_{ij} of an opinion score $R_{j,i}$ being anomalous increases, the proposed reliability measure accurately indicates a decrease in its reliability. Moreover, it is noteworthy that the CI of the reliability decreases in size as the probability of the opinion score being anomalous increases.

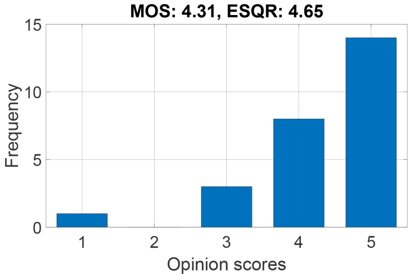


Fig. 8. Quality estimates of ESQR vs. other methods. Opinion scores distribution for a stimulus with high discrepancy between Q (ESQR) and the mean opinion score (MOS).

This suggests that the proposed measures exhibit higher confidence in the estimation of reliability when the opinion is almost certainly affected by an anomaly.

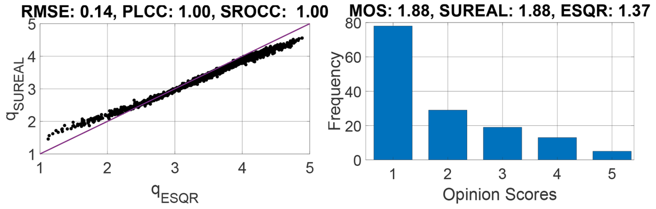
Additionally, an interesting observation is that biased opinion scores maintain greater reliability compared to spam scores. This indicates that the proposed measure of reliability can discern that biased opinion scores contain more valuable information compared to spam scores, which are essentially noise. In summary, the simulation results affirm the effectiveness of the proposed non-parametric measure of reliability in correctly identifying and quantifying the impact of anomalies on opinion scores.

G. *Quality Estimates of ESQR vs. Other Methods*

We evaluate the similarity between the subjective quality recovered by ESQR and that of other state-of-the-art quality estimation algorithms on the six considered datasets. ESQR estimates generally align with prior methods, deviating only in specific cases where assumptions are potentially violated. The smallest Pearson correlation found between ESQR and other methods is 0.996 (0.994 for Spearman correlation), indicating very high consistency on average. The RMSE further confirms these results, with a maximum value of 0.167, notably small on a 5-level quality scale.

It is instructive to analyze cases where the quality estimated by ESQR differs significantly from that of alternative methods. An example of stimulus where the output of ESQR deviates significantly from the MOS (difference = 0.34) is the one whose distribution of scores is showed in Fig. 8. In fact, while 14 subjects out of 26 deemed that the quality of that stimulus was excellent and scored it with a 5, there is one subject that found the quality bad and gave a 1 as opinion score. The MOS attributes to this opinion score the same importance that is attributed to the other opinion scores; this yields a MOS = 4.31. The proposed ESQR algorithm instead under-weights that potentially noisy low opinion score and recovers a larger subjective quality (4.65).

To study significant differences between ESQR and the popular Netflix SUREAL, we compare the quality estimates by the two methods for the large-scale dataset MoviesLens-1 M through the scatter plot in Fig. 9(a). We observe that the output of ESQR mostly differs from that of the Netflix SUREAL software at the extremes of the quality scale, i.e., where the



(a) Netflix SUREAL vs ESQR (b) Stimulus with largest difference

Fig. 9. Quality estimates of ESQR vs. other methods. Comparing the output of the ESQR algorithm to the one of the Netflix SUREAL software on the MovieLens-1 M dataset. On the right, the distribution of opinion scores of the stimulus for which the outputs of the two approaches differ the most.

TABLE IV
Ablation Study on the ESQR Algorithm

Index	ESQR	Without Pairwise Correlation	Exp value computed from $\hat{p}_{V_i}$
Rob. to Noise (RMSE)	**0.06**	**0.06**	0.10
Rob. to Spammer Annot. (RMSE)	0.06	0.12	**0.04**
Average Size of CIs	**0.36**	0.38	0.51
Per Stimulus Reliability	**yes**	**yes**	no

"The best performance for each testing condition is highlighted in bold".

distribution of the opinion scores is typically strongly asymmetric. In these cases, the Gaussianity assumption made by Netflix SUREAL concerning the quality distribution is not met, leading the method to introduce substantial bias in the estimates. We observed a similar behavior for KoNViD-1 k. Fig. 9(b) reports the score distribution for the stimulus having the largest estimated quality difference between the ESQR algorithm and Netflix SUREAL. The Netflix SUREAL software recovered a subjective quality equal to the MOS. The ESQR algorithm instead recovered a subjective quality that is significantly lower than the MOS. The difference, 0.51, represents more than 10% of the whole quality scale. This confirms that, despite most of the time ESQR provides estimates coherent with previous methods, there are stimuli for which different quality recovery approaches strongly disagree.

H. Ablation Studies

In our ESQR algorithm ablation study we examined two aspects: i) removing pairwise correlation, thus directly using the observed distribution $\bar{p}_{V_i}$ (11) instead of the estimate $\hat{p}_{V_i}$ (13) that involves the weights ϵ_{ij}; ii) using the expected value computed from $\hat{p}_{V_i}$ as the subjective quality estimate. Results are shown in Table IV. We focus only on the Netflix Public dataset since similar conclusions were drawn from other datasets. To evaluate the robustness against noise and spammer annotators, we computed the average RMSE across the noisy scenarios in Figs. 2 and 3.

As it can be noticed from Table IV, without pairwise correlation, the algorithm is less robust to spammer annotators, and by using only the expected value computed from $\hat{p}_{V_i}$, it is less robust to noise and computes a quality estimate with larger CIs and thus prone to more uncertainty. The full ESQR algorithm demonstrates superior performance balance, highlighting the importance of all introduced elements.

It is worth noting from (9) that our introduced reliability measure plays a crucial role in our ESQR algorithm. Removing this measure, or equivalently assuming that opinion scores are equally reliable, would reduce ESQR to the MOS. However, throughout all experiments presented in this paper, ESQR consistently outperformed the MOS. This superiority is manifested in various aspects: i) ESQR demonstrates the capability to estimate subjective quality with lower uncertainty than the MOS, as evidenced by the results presented in Table I. ii) ESQR provides better predictions of confidence intervals, as indicated in Table II. iii) ESQR exhibits higher robustness to noise, as highlighted by the results in Section V-D. These results clearly demonstrate that the enhanced accuracy of ESQR stems from the weighting of opinion scores by our proposed reliability measure. Thus it is a fundamental piece of the algorithm.

I. Computational Time Analysis

Each method was executed 30 times on each dataset, and the average computational time was recorded. These experiments were conducted using MATLAB on a computer equipped with a 2.6 GHz 6-Core Intel Core i7 processor and 16 GB of RAM.

Excluding RMLE, all approaches processed the small-scale datasets (VQEG-HD1, VQEG-HD3, VQEG-HD5, and Netflix Public) in less than 4 ms on average. Specifically, Netflix SUREAL and ZREC took less than 1 ms each, while ESQR required just over 3 ms. A notable increase in computational time (up to 70 seconds) was observed with RMLE on small-scale datasets. This highlights the computational demands of parametric methods in estimating optimal parameter values. The efficient processing by SUREAL and ZREC is attributed to the approximation of the parameter estimation process through an iterative procedure and utilization of statistical moments, respectively.

Finally, it is interesting to note that, despite being slightly slower than ZREC and the Netflix SUREAL software, ESQR completed the recovery of the subjective quality on the MovieLens-1 M dataset, involving up to 1 million ratings, in no more than 12 seconds. This clearly suggests that the efficiency of ESQR is not questionable for practical exigencies.

VI. Conclusion

In this paper, we introduce ESQR, a novel Entropy-based Subjective Quality Recovery algorithm to estimate subjective media quality from noisy opinion scores. The primary idea behind our approach is to treat quality estimation as a non-parametric problem, diverging from the prevalent practice in the literature that involves modeling scoring behavior through predefined and often simplistic distributions. Specifically, we establish a reliability measure for each stimulus capturing the degree of surprise that a given score brings compared to the overall score distribution. We then utilize this measure to weigh the contribution of individual opinion scores to the overall quality of a stimulus.

When comparing ESQR to five state-of-the-art quality recovery methods across six diverse datasets, our results indicate that:

i) ESQR produces subjective quality estimates characterized by reduced uncertainty; ii) ESQR demonstrates superior robustness to noise compared to other methods; and iii) ESQR maintains its accuracy across a broader range of applications and datasets.

Future work will explore a refinement of the introduced reliability measure to explicitly consider the ordinal nature of quality scales, as entropy-based approaches may overlook this essential aspect.

Appendix A
Asymptotic Distribution of Quality Estimator

Consider a stimulus that has been scored independently by M subjects. The subject j has given a score $1 \leq R_j \leq K < \infty$ drawn from a distribution p_{R_j}. Based on the scores $\{R_1, \ldots, R_M\}$ a quality recovery algorithm is proposed as a weighted average:

$$Q(M) = \frac{\sum_{j=1}^{M} h(R_j) R_j}{\sum_{l=1}^{M} h(R_l)}. \tag{21}$$

In the ESQR algorithm we would have $h(R_j) = -\log(\hat{p}_V(R_j))^{-1}$ where $\hat{p}_V$ is an approximation to the scoring distribution. In this section, to simplify our theorical analysis we assume that we are in fact using the true scoring distribution p_V. Other algorithms (see (19)) have the same structure so our analysis can also be extended to those cases.

Proposition 2: If there exists $c > 0$ such that $|h(R_j)| \leq c$ for all j, and as $M \to \infty$ for any $q \in [1,\ K]$ we have that:

$$\sum_{j=1}^{M} \mathrm{var}\left[h(R_j)(R_j - q)\right] \to \infty, \tag{22}$$

then the asymptotic distribution of Q is:

$$\lim_{M \to \infty} \mathbb{P}(Q \leq q) = \Phi\left(-\frac{\mu_q}{\sigma_q}\right), \tag{23}$$

where Φ is the distribution function of a standard normal random variable and:

$$\mu_q = \sum_{j=1}^{M} \mathbb{E}\left[h(R_j)(R_j - q)\right], \tag{24}$$

$$\sigma_q^2 = \sum_{q=1}^{M} \mathrm{var}\left[h(R_j)(R_j - q)\right]. \tag{25}$$

Proof: Using (21) we may write the distribution of Q as:

$$F_Q(q) = \mathbb{P}(Q \leq q) = \mathbb{P}\left(\sum_{j=1}^{M} h(R_j)(R_j - q) \leq 0\right). \tag{26}$$

Since $|h(R_j)| \leq c$, for some c, and $q \in [1,\ K]$ then the random variables in the summation are uniformly bounded for all j as $|h(R_j)(R_j - q)| \leq c(K - 1)$. Under (22) we can now apply Lindeberg's central limit theorem (CLT) [52, Example 27.4] to obtain the desired result. □

Notice that event though we have a closed form approximation for the distribution of Q using the the CLT we cannot guarantee that Q will be normally distributed since μ_q and σ_q are functions of q. The approximation will yield a normal distribution if and only if μ_q/σ_q is a linear function of q. By looking at (24) we see that μ_q is indeed a linear function of q. Then we have the following corollary:

Corollary 1: Under the hypotheses of Proposition 2, Q is asymptotically normal if and only if (25) is independent of q.

It is clear that a good approximation will be retained as long as μ_q/σ_q is approximately linear in q where Φ changes more rapidly. We now perform some numerical simulations to study whether the quality estimate can be assumed to be normal. In order to do this, we need to test two things:

T1) Using the CLT is a good approximation for F_Q for the number M of subjects typically considered. This would validate that (23) is a good approximation for practical finite values of M.

T2) The argument of (23) is a linear function of q where Φ changes rapidly, which, together with T1) would validate that the estimator is approximately normal in practice.

For the tests, we perform simulation with $M = 24$ subjects. We consider a model very similar to that of Section V-B. We assume that each subject independently rates the same stimulus, giving a score on $\{1, 2, 3, 4, 5\}$. The scores is given according to the true distribution p_V with probability $1 - p_e$, and a uniform score with probability p_e. The probability p_e, different for each subject, is obtained as a uniform random variable on $(0, 0.05)$. The distribution p_V is obtained by discretizing continuous distributions, namely:

- A normal random variable of mean x_e and deviation a, where x_e is drawn from a continuous uniform random variable on (1,5) and $a = 0.2 \times (-x_e^2 + 6x_e - 5)$ [33]. The discretization is done considering the points $\{1.5, 2.5, 3.5, 4.5\}$ of the normal variable.
- A beta random variable with parameters a and b drawn as independent continuous uniform variables on (1,10). The discretization is done by dividing the support (0,1) into 5 consecutive equally spaced intervals.

To test the validity of T1) and T2) we proceed as follows:

- Let $\mathbf{q}$ be a uniform grid of n_d points in $[1, 5]$, that is:

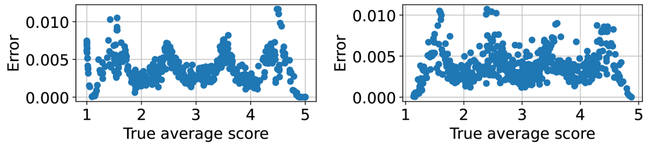
$$\mathbf{q} = \left\{\frac{4i + (n_d - 5)}{n_d - 1} : i = 1, \ldots, n_d\right\}. \tag{27}$$

These points are were F_Q will be estimated.

- Choose an input distribution for V, normal or beta.
- Do the following experiment n_p times:

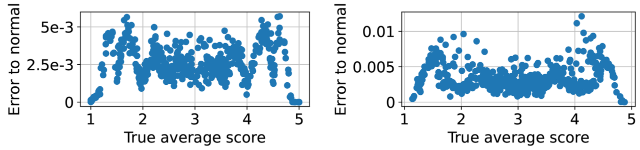
1) Choose the parameters θ for the true score, with $\theta = (x_e, a)$ for normal and $\theta = (a, b)$ for beta.
2) For each $q \in \mathbf{q}$ obtain estimates $\hat{\mu}_q$ and $\hat{\sigma}_q$ of (24) and (25), respectively, through samples averages using $n_s = 5000$ independent realizations of the scores of the M subjects. Then obtain the estimate of the CDF of Q using the CLT as: $\hat{F}_{Q,\text{CLT}}(q) = \Phi(\frac{\hat{\mu}_q}{\hat{\sigma}_q})$.
3) For each $q \in \mathbf{q}$ obtain an independent estimate of the distribution of Q using n_s independent realizations of the scores as:

$$\hat{F}_{Q,\text{emp}}(q) = \frac{\#\text{ Samples of } Q \leq q}{n_s}. \tag{28}$$



(a) Normal. Average L1 error is 0.004. (b) Beta. Average L1 error is 0.004.

Fig. 10. **Test T1.** Scatter plot of the L^1 error between (23) for finite M and the empirical estimation of Q as a function of the true average score.



(a) Normal. Average L1 error is 0.003. (b) Beta. Average L1 error is 0.003.

Fig. 11. **Test T2.** Scatter plot of the L^1 error between (23) for finite M and a normal distribution with the same mean and variance as Q, as a function of the true average score.

Also, compute the sample mean $\hat{\mu}_{Q,\text{emp}}$ and sample variance $\hat{\sigma}^2_{Q,\text{emp}}$.

4) *Computations for T1:* $\hat{F}_{Q,\text{CLT}}$ uses Proposition 2 (the CLT) to approximate the distribution of Q, while $\hat{F}_{Q,\text{emp}}$ does not make any modeling assumptions. If the CLT is a good approximation, then both estimators should give similar values, which validates T1). To check this, we estimate this error using the L^1 norm of the error between the two estimators by using the discrete samples in $\mathbf{q}$. To do this we compute the error: $e_i = |\hat{F}_{Q,\text{CLT}}(q_i) - \hat{F}_{Q,\text{emp}}(q_i)|$ where $q_i = \frac{4i+(n_q-5)}{n_q-1}$ and estimate the L^1 error between the two estimators using the trapezoidal rule. where Δ is the spacing between two values of q.

5) *Computations for T2:* validating T1, this does not mean that Q is approximately normal. To verify this we compare $\hat{F}_{Q,\text{CLT}}$ with the distribution of a normal with mean $\hat{\mu}_{Q,\text{emp}}$ and variance $\hat{\sigma}^2_{Q,\text{emp}}$ by computing the L^1 in the same manner as with $\hat{F}_{Q,\text{emp}}$.

After the n_p repetitions we have computed the n_p estimates of $\hat{F}_{Q,\text{CLT}}$, $\hat{F}_{Q,\text{emp}}$, and the distribution of a normal with mean $\hat{\mu}_{Q,\text{emp}}$ and variance $\hat{\sigma}^2_{Q,\text{emp}}$, for different parameters of the input distribution. We also computed the L^1 error between the $\hat{F}_{Q,\text{CLT}}$ and the other two estimates. If both errors are small then T1 and T2 are validated, which means that, at least for the proposed distributions, the CLT is a good approximation for values as small as $M = 24$ subjects and that Q is approximately normal.

In Fig. 10 we can see the scatter plot of the L1 error of the true distribution of Q and $\hat{F}_{Q,\text{CLT}}$, the approximation (23) using the CLT for finite M, for $n_p = 500$ repetitions and the three score distributions. We see that the total worst case error is very small, below 0.012, for all the possible true scores. In Fig. 11 we see the scatter plot of the L1 error between $\hat{F}_{Q,\text{CLT}}$ and a normal distribution with the same mean and variance as Q. Again for all the results the worst error is very small, around 0.006 for Gaussian and 0.012 for the Beta.

References

[1] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Proc. IEEE Data Compression Conf.*, 2017, pp. 52–61.

[2] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests-beyond subjects' MOS," *IEEE Trans. Multimedia*, vol. 23, pp. 2505–2519, 2021.

[3] 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 *Proc. 28th Int. Conf. Multimedia.* , 2020, pp. 3339–3347.

[4] L. F. Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Modeling subject scoring behaviors in subjective experiments based on a discrete quality scale," *IEEE Trans. Multimedia*, 2024, early access, Mar. 27, 2024, doi: 10.1109/TMM.2024.3382483.

[5] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-T Standard Rec. BT.500, ITU Radiocommunication Sector, Geneva, Switzerland, 2019.

[6] Z. Li, C. G. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," in *Proc. Int. Symp. Electron. Imag.*, 2020, pp. 1–14.

[7] L. F. Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings," in *Proc. IEEE 14th Int. Conf. Qual. Multimedia Experience*, 2022, pp. 1–4.

[8] L. F. Tiotsop, A. Servetti, and E. Masala, "A scoring model considering the variability of subjects' characteristics in subjective experiments," in *Proc. IEEE 15th Int. Conf. Qual. Multimedia Experience*, 2023, pp. 43–48.

[9] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Trans. Multimedia*, vol. 17, pp. 2210–2224, 2015.

[10] ITU-T, "Subjective evaluation of media quality using a crowdsourcing approach," ITU-T, Geneva, Switzerland, Tech. Rep. ITU-T PSTR-CROWDS, 2018.

[11] H. Lin et al., "Large-scale crowdsourced subjective assessment of picturewise just noticeable difference," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5859–5873, Sep. 2022.

[12] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10 K: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[13] U. Reiter et al., "Factors influencing quality of experience," in *Quality of Experience*, S. Möller and A. Raake, Eds., Cham, Switzerland: Springer, 2014, pp. 55–72.

[14] D. Schwarz, G. Lemaitre, M. Aramaki, and R. Kronland-Martinet, "Effects of test duration in subjective listening tests," in *Proc. Int. Comput. Music Conf.*, 2016, pp. 515–519.

[15] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, Jul. 2015.

[16] C. Deng, L. Ma, W. Lin, and K. N. Ngan, Eds., *Visual Signal Quality Assessment: Quality of Experience (QoE)*, 1st ed. Cham, Switzerland: Springer, 2015.

[17] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, Mar. 2011.

[18] J.-S. Lee, L. Goldmann, and T. Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia Tools Appl.*, vol. 67, pp. 31–48, 2013.

[19] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, Dec. 2012.

[20] P. Kortum and M. Sullivan, "The effect of content desirability on subjective video quality ratings," *Hum. Factors*, vol. 52, no. 1, pp. 105–118, Feb. 2010.

[21] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Online HodgeRank on random graphs for crowdsourceable QoE evaluation," *IEEE Trans. Multimedia*, vol. 16, pp. 373–386, 2014.

[22] Q. Xu et al., "Exploring outliers in crowdsourced ranking for QoE," in *Proc. 25th Int. Conf. Multimedia*, 2017, pp. 1540–1548.

[23] J. Li, R. Mantiuk, J. Wang, S. Ling, and P. Le Callet, "Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3479–3489.

[24] E. Zerman, G. Valenzise, and A. Smolic, "Analysing the impact of cross-content pairs on pairwise comparison scaling," in *Proc. IEEE 11th Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–6.

[25] M. Perez-Ortiz et al., "From pairwise comparisons and rating to a unified quality scale," *IEEE Trans. Image Process.*, vol. 29, pp. 1139–1151, 2020.

[26] E. Zerman, V. Hulusic, G. Valenzise, R. Mantiuk, and F. Dufaux, "The relation between MOS and pairwise comparisons and the importance of cross-content comparisons," in *Proc. Hum. Vis. Electron. Imag. Conf., IS&T Int. Symp. Electron. Imag.*, 2018, pp. 1–6.
[27] *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T Rec. P.910, ITU-T Recommendations, Geneva, Switzerland, 2022.
[28] *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in any Environment*, ITU-T Rec. P.913, ITU-T Recommendations, Geneva, Switzerland, 2021.
[29] R. B. Millar, *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. Hoboken, NJ, USA: Wiley, 2011.
[30] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.
[31] J. Zhu, A. Ak, P. Le Callet, S. Sethuraman, and K. Rahul, "ZREC: Robust recovery of mean and percentile opinion scores," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 2630–2634.
[32] Netflix, "The Sureal software," 2022. Accessed: Dec. 16, 2022. [Online]. Available: https://github.com/Netflix/sureal
[33] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in *Proc. IEEE 3rd Int. Workshop Qual. Multimedia Experience*, 2011, pp. 131–136.
[34] V. C. Raykar et al., "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, no. 43, pp. 1297–1322, 2010. [Online]. Available: http://jmlr.org/papers/v11/raykar10a.html
[35] J. Zhang, V. S. Sheng, T. Li, and X. Wu, "Improving crowdsourced label quality using noise correction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1675–1688, May 2018.
[36] S. Li, X. Xia, S. Ge, and T. Liu, "Selective-supervised contrastive learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 316–325. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00041
[37] S. Li, T. Liu, J. Tan, D. Zeng, and S. Ge, "Trustable co-label learning from multiple noisy annotators," *IEEE Trans. Multimedia*, vol. 25, pp. 1045–1057, 2023, doi: 10.1109/TMM.2021.3137752.
[38] S. Li et al., "Coupled-view deep classifier learning from multiple noisy annotators," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4667–4674. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5898
[39] D. Cheng et al., "Class-dependent label-noise learning with cycle-consistency regularization," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, pp. 11104–11116. [Online]. Available: https://openreview.net/forum?id=IvnoGKQuXi
[40] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 03, pp. 447–461, Mar. 2016.
[41] S. Li, X. Xia, H. Zhang, Y. Zhan, S. Ge, and T. Liu, "Estimating noise transition matrix with label correlations for noisy multi-label learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24184–24198.
[42] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 68–83.
[43] R. Wang, T. Liu, and D. Tao, "Multiclass learning with partially corrupted labels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2568–2580, Jun. 2018.
[44] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, "Repeated labeling using multiple noisy labelers," *Data Mining Knowl. Discov.*, vol. 28, no. 2, pp. 402–441, Mar. 2014, doi: 10.1007/s10618-013-0306-1.
[45] C. Kang, G. Valenzise, and F. Dufaux, "Predicting subjectivity in image aesthetics assessment," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process.*, 2019, pp. 1–6.
[46] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Hoboken, NJ, USA: Wiley, 2006.
[47] N. C. Silver and W. P. Dunlap, "Averaging correlation coefficients: Should Fisher's Z transformation be used?," *J. Appl. Psychol.*, vol. 72, no. 1, 1987, Art. no. 146.
[48] B. Ćmiel, J. Nawała, L. Janowski, and K. Rusek, "Generalised score distribution: Underdispersed continuation of the beta-binomial distribution," *Stat. Papers*, vol. 65, no. 1, pp. 381–413, 2024.
[49] VQEG, "Report on the validation of video quality models for high definition video content (v. 2.0)," 2010. Accessed: Dec. 16, 2022. [Online]. Available: https://bit.ly/2Z7GWDI
[50] V. Hosu et al., "The konstanz natural video database (KoNViD-1 k)," in *Proc. IEEE 9th Int. Conf. Qual. Multimedia Experience*, 2017, pp. 1–6.
[51] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interactive Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2015.
[52] P. Billingsley, *Probability and Measure*, 2nd ed. Hoboken, NJ, USA: Wiley, 1986.

Andrés Altieri received the Ingeniero Electrónico degree (Hons.) in electronics engineering from the University of Buenos Aires (UBA), Buenos Aires, Argentina, in 2009, and the joint Ph.D. degree from UBA and École Supérieure d'Électricité (currently CentraleSupélec), Gir-sur-Yvette, France, in 2014. He is currently an Assistant Professor with the School of Engineering, UBA, and a Researcher with the National Council for Scientific and Technological Research (CONICET), Buenos Aires. Since 2022, he has been a Researcher with the Laboratoire des Signaux et Systèmes, CentraleSupélec, CNRS, Université Paris-Saclay, Gif-sur-Yvette. His research interests include statistical signal processing and learning, microwave hardware design, and wireless sensing.

Lohic Fotio Tiotsop (Member, IEEE) received the M.Sc. degree in mathematical engineering and the Ph.D. degree in control and computer engineering from Politecnico di Torino, Torino, Italy, in 2017 and 2021, respectively. He is currently a Postdoctoral Researcher with the Control and Computer Engineering Department, Politecnico di Torino. His primary research interests include statistical models, machine learning, and deep learning-based approaches, specifically delving into media quality assessment.

Giuseppe Valenzise (Senior Member, IEEE) received the Ph.D. degree in information technology with the Politecnico di Milano, Torino, Italy, in 2011. He is currently a CNRS Researcher with Laboratoire des Signaux et Systèmes (L2S), Université Paris-Saclay, CentraleSupélec, Gir-sur-Yvette, France, where he is the Head of the Multimedia and Networking Team. In 2012, he joined the French Centre National de la Recherche Scientifique (CNRS) as a permanent Researcher, first with the Laboratoire Traitement et Communication de l'Information (LTCI) Telecom Paristech, and in 2016 with L2S. He is the coauthor of more than 100 research publications and of several award-winning papers. His research interests include different fields of image and video processing, traditional and learning-based image and video compression, immersive video (light fields, point clouds), image/video quality assessment, high dynamic range imaging, and applications of machine learning to image and video analysis. He was the recipient of the EURASIP Early Career Award in 2018 for "significant contributions to video coding and analysis". He was/is an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing, *Elsevier Signal Processing: Image communication*. He is the Chair of the MMSP Technical Committee of the IEEE Signal Processing Society for the term 2024–2025, and he was a member of the Technical Area Committee on Visual Information Processing of EURASIP from 2018 to 2023.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement