Quality of Experience in Video Streaming: Status Quo, Pitfalls, and Guidelines

Leonardo Peroni
IMDEA Networks Institute and UC3M
Madrid, Spain
leonardo.peroni@imdea.org

Sergey Gorinsky
IMDEA Networks Institute
Madrid, Spain
sergey.gorinsky@imdea.org

***Abstract*—Quality of experience (QoE) becomes both the holy grail and a free-for-all in adaptive bitrate (ABR) video streaming. On the one hand, the design, operation, and evaluation of ABR algorithms increasingly rely on QoE. On the other hand, QoE frequently receives only cursory attention in this supporting role, with many of its important aspects treated with insufficient care. As a complex subjective notion, QoE is directly measurable through subjective tests, which incur evident overhead. While an objective QoE model represents a scalable automated means for QoE assessment, QoE models proliferate without consensus on their goodness due to numerous influence factors, construction methods, and usages. The model proliferation creates a false impression that proposing a new QoE model without a proper validation is acceptable. Because the multifaceted QoE problem involves separable and often separated tasks of test conducting, model building, and model using, this separation of concerns causes additional complications. By leveraging two large real datasets of individual QoE perception, this paper reviews the status quo in QoE, identifies various pitfalls, and offers guidelines for test conducting, model building, and model using, so as to foster high standards in future work on QoE in ABR video streaming.**

***Index Terms*—Video streaming; quality of experience; subjective test; QoE model; scoring scale; interface design; experience selection; value interpretability; range capping; evaluation metric; ABR algorithm.**

I. Introduction

Quality of experience (QoE) plays an important role in the design, operation, and evaluation of networked computer systems that serve humans. Qualinet, a European Cooperation in Science and Technology Action, provides a two-sentence definition for QoE [1]. This definition, endorsed by the International Telecommunication Union (ITU) [2] and widely cited in general, distinguishes two pertinent aspects of QoE. First, QoE captures the overall satisfaction of a user with an application as perceived by this user, i.e., QoE is a subjective personal notion. Second, QoE depends on the user's current state that has multiple dimensions, e.g., network connectivity, device type, and application content.

This paper studies QoE in adaptive bitrate (ABR) video streaming [3], an application that heavily dominates the Internet traffic [4], [5]. The origin server of an ABR streaming session partitions the video into chunks and encodes every chunk into multiple representations in the form of bitrate-resolution pairs. The ABR algorithm of each client independently requests a representation for the next chunk in order to handle varying network conditions and clashing performance objectives. When the requested bitrate is too high, the chunk arrives too late for uninterrupted playback, and the resulting stall of the video at the client degrades QoE. On the other hand, requesting a low bitrate reduces the video quality and thereby hampers QoE too. Although the terminology and discussion in this paper are for ABR streaming, we consider the paper's general takeaways as being relevant to QoE in other kinds of networked computer systems.

While appealing as a basis for user-centered system design, operation, and evaluation, QoE raises a variety of practical complications. In particular, QoE subjectivity implies that direct assessment of QoE involves subjective tests where human raters provide scores for experiences presented to them. However, lab-based subjective assessments consume significant amounts of time and effort, and online crowdsourcing alternatives mitigate the overhead concerns only to some extent [6], [7].

The overhead of subjective tests fuels the emergence and wide spread of QoE models. A QoE model automatically derives QoE from objective influence factors (IFs), such as stall duration and bitrate changes across consecutive chunks [8]. Traditionally, the construction of a QoE model presents a series of experiences to a group of raters, averages the raters' individual scores to compute the mean opinion score (MOS) of each experience, and approximates QoE as a function mapping the considered IFs to MOS. The advantage of the traditional QoE modeling is that only a relatively small group of raters participates in subjective tests whereas the constructed QoE model automates QoE assessment for all users of the application without imposing any subjective-test overhead on a huge majority of them.

Despite offering the scalable automated support for QoE assessment, QoE models spawn new difficulties. Human perception of video is complex, and many IFs of different kinds are pertinent to QoE [9]–[13]. Besides, practical considerations necessitate that the IFs of a QoE model are measurable by the entity that uses the QoE model, with these measurements being sufficiently accurate and incurring only low overhead. For example, although research indicates promise of electroencephalographic signals as IFs of QoE [14], a streaming provider is unlikely to deploy a large-scale application that attaches electrodes to the users' scalps. Instead, it is typical for an application provider to directly measure stall duration and estimate available network bandwidth based on throughput observations in the client. QoE models also diverge with respect to the approximation function that maps the considered IFs to QoE. For instance, closed-form expressions and learning-based approaches are both common in QoE modeling. Consequently, there exist a large number of diverse QoE models.

The diversity of QoE models also arises due to different usages of the models. Timing and accuracy considerations might necessitate different models for design, operation, and evaluation of systems. In particular, a complex QoE model might be suitable for offline design or evaluation but not for real-time operation of an ABR algorithm. For example, whereas the peak signal-to-noise ratio (PSNR) [15] and video multimethod assessment fusion (VMAF) [16] are metrics of video quality that dramatically differ in their computational requirements, [17] relies on PSNR to predict video quality during live streaming and leverages VMAF to evaluate the actually achieved video quality.

Besides, the multifaceted QoE problem involves separable tasks such as test conducting, model building, and model using. *Test conducting* performs subjective tests. *Model building* constructs QoE models based on subjective scores. *Model using* utilizes QoE models in system design, operation, or evaluation. A single work might handle multiple tasks. For instance, iQoE [18] both conducts subjective assessments and constructs personalized QoE models. Sensei [19] and Ruyi [20] address all three tasks of test conducting, model building, and model using. ARTEMIS [17] neither builds nor validates a QoE model and instead uses an existing QoE model to evaluate its proposal that dynamically configures the bitrate ladder of a live ABR streaming session. The separation of concerns in dealing with QoE has both positives and negatives. On the one hand, the focus on a single task enables its more thorough execution. On the other hand, the limited outlook might derail the overall effort, e.g., when an ABR algorithm uses a QoE model validated for dissimilar settings.

The importance, complexity, and separation of concerns put QoE in a precarious position. The widely recognized importance of QoE creates expectations to consider QoE in ABR video streaming, at least for evaluation if not for design and operation. However, QoE complexity makes comprehensive treatment of QoE difficult. Furthermore, the diversity of existing QoE models creates a false impression that one may easily introduce a new QoE model without a proper validation. Hence, QoE becomes both the holy grail and a free-for-all.

In this paper, we review the current landscape of QoE in ABR video streaming and zoom in on a number of areas including: (a) scoring scale, interface design, and experience selection for subjective tests, (b) validation, value interpretability, and capping of the value range in QoE modeling, (c) mismatch between usage and construction of QoE models, (d) evaluation of QoE models via correlation vs. error metrics, and (e) QoE evaluation of ABR algorithms. We identify problems afflicting these areas and offer advice on how to rectify the situation. Our methodology leverages real data and arranges the advice in accordance with the classification of QoE-related tasks into test conducting, model building, and model using. In recommending the good practices for subjective assessments, construction and usage of QoE models, our overarching aspiration is to foster high standards in future work on QoE in ABR streaming. While we expect the increased awareness and good practices to be the most valuable for newcomers to the field, this paper serves as a wake-up call for the entire community to acknowledge and address the identified problems.

II. Background

QoE has its roots in quality of service (QoS), an earlier notion from packet-switched computer networking. QoS characterizes network performance via such metrics as the transmission rate, packet loss, end-to-end delay, and delay jitter provided to applications [21]. Two main features differentiate QoE from QoS. First, QoE shifts the focus from objective system performance to the user's subjective perception of the performance. Second, while QoS is rather an umbrella term for multiple metrics, QoE constitutes a holistic concept capturing the user's overall satisfaction with the application. The evolution from network-centered QoS to user-centered QoE not only fulfills the interests and needs of application providers but also is relevant to network operators. For example, a network operator might utilize a QoE model as a basis for allocation of link capacities to video streams [22].

In subjective QoE tests of ABR video streaming, an experience refers to a sequence of chunks played back by the client to a rater who provides a score for the experience. When a subjective test collects scores for a series of experiences to support construction of a QoE model, the test also records the IF values of each rated experience. To keep the load on the raters manageable, the series of experiences should be relatively short. [23], [24], [25], and, to a smaller extent, [26] select the experiences and their IF values to be representative of real-world settings.

The scoring scale is an important element of subjective testing methodologies, including those standardized by ITU [27]. For instance, absolute category rating (ACR) is a popular method with a five-level scale where integers from 1 to 5 constitute bad, poor, fair, good, and excellent levels [28]. Another common scale consists of 100 levels where level ranges 1-20, 21-40, 41-60, 61-80, and 81-100 correspond to bad, poor, fair, good, and excellent QoE, respectively [18], [23], [24], [29]. While such discrete absolute scales are the most typical, alternative testing methods employ continuous scales for scoring an experience, assess QoE degradation rather than QoE itself, or perform pairwise comparison of experiences [27], [30]. According to [31] and [32], usage of continuous vs. discrete scales results in no significant statistical differences in QoE assessment.

Building a QoE model based on the experiences' scores and IF values has many methods of different kinds at its disposal. Although classification techniques seem a natural fit for modeling of discrete QoE scores, regression methods dominate QoE modeling. In this paper, we consider 10 existing QoE models and, for brevity, refer to them with the following single-letter labels: B [33], G [34], R [35], S [36], V [37], N [38], F [39], A [40], P [41], and L [42]. The first six of these QoE models rely on regression with similar linear target functions and account for video quality with a different IF. The target function of QoE model F is exponential. The construction of QoE models A, P, and L relies on machine learning and, specifically and respectively, on support vector regression, random forest, and long short-term memory.

When constructed, a QoE model avails itself to various usage in system design, operation, and evaluation. MPC [33] makes ABR decisions via model predictive control based on QoE model B (as labeled above). Pensieve [43] is an ABR algorithm that uses QoE model B as the optimization objective in actor-critic reinforcement learning. Although BBA [44] and ThroughputRule (TR) [45] do not rely on any QoE model in their design or operation, usage of QoE models to evaluate QoE performance of such ABR algorithms is common as well.

III. Methodology

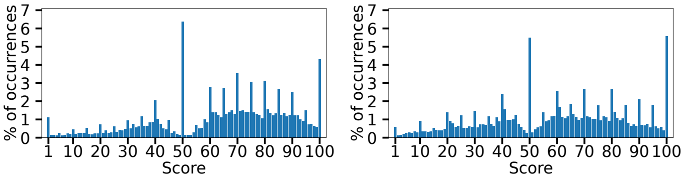
The nine subsequent sections identify and examine problems in dealing with QoE by progressively covering the tasks of test conducting, model building, and model using. In the process, we cite additional problem-specific related work and, when needed, utilize the aforementioned QoE models and ABR algorithms. The analysis in each of these sections offers advice on addressing the examined problem. Because our analyses heavily leverage two large real datasets of QoE perception by individual raters on the 1-100 scoring scale, we now describe these Waterloo-IV and iQoE datasets in more detail.

Waterloo-IV [46] is a dataset with 43,650 individual scores from lab experiments with 92 raters aged between 18 and 38 years old, with 29, 32, and 31 of the raters using phone, high definition television (HDTV), and ultra HDTV devices, respectively. The presented experiences span all combinations of two codecs, nine network traces, and five ABR algorithms. 13 IFs characterize every chunk, which has the duration of 4 s. Each experience consists of seven chunks, i.e., the playback of an experience without stalls takes 28 s.

iQoE [47] refers to a dataset with 14,400 individual scores from online subjective tests with 120 raters aged between 20 and 63 years old. Among the raters who disclose their viewing device, six and 110 raters claim using a phone and personal computer, respectively. The iQoE dataset contains 1,000 experiences generated via simulations in Park [48] by utilizing one codec, 102 network traces, and three ABR algorithms. Each experience contains four chunks characterized by 10 IFs. Because the chunk duration is set to 2 s, each experience plays back for 8 s without stalls.

IV. Scoring Scale in Subjective Assessments

Conducting a subjective test involves selecting a scale for scoring of experiences. The Waterloo-IV and iQoE datasets described in Section III employ the 1-100 scale. Compared to the five-level ACR scale, the 1-100 scale gives the raters an opportunity to express their QoE perception with a finer granularity, which might make the QoE assessment more accurate. On the other hand, the 100 levels increase the raters' uncertainty about which specific level to choose, and the increased cognitive load on the raters might degrade the assessment accuracy due to hasty or careless decisions.



(a) Waterloo-IV HDTV

(b) iQoE

Fig. 1: Distributions of the individual scores in the datasets.

To analyze how the number of scale levels affects QoE rating, we consider the distributions of individual scores in the Waterloo-IV and iQoE datasets. Because the experiences chosen by the datasets deliberately cover the entire 1-100 scale to support construction of accurate QoE models, we expect the popularity of the individual scores across the scale to be smooth if not uniform. With the uniform distribution, the popularity of each score would be 1%.

For the 32 HDTV raters of the Waterloo-IV dataset, Figure 1a depicts the distribution of the score popularity that is neither uniform nor close to being smooth. Instead, there is a small number of scores that spike in popularity compared to the adjacent scores. In particular, the score of 50 dominates by grabbing 6.37% of all score occurrences and apparently drawing attention to itself at the expense of the other scores in the 41-59 range. The next nine popular scores, in decreasing order of popularity, are 100, 70, 80, 75, 60, 65, 85, 90, and 40 that capture 4.30%, 3.55%, 3.14%, 3.07%, 2.77%, 2.70%, 2.69%, 2.49%, and 2.05% of all score occurrences, respectively. The results suggest that, in agreement with prototype theory [49], the raters form their own new categories of scores where the prototype of each category is either a score divisible by five or the lowest score of 1. When presented with an experience, a rater determines a matching new category and reports the category prototype as the score for the experience.

Figure 1b plots the 14,400 individual scores in the iQoE dataset. Despite conducting the tests in online rather than lab settings, the qualitative results are remarkably consistent with those for Waterloo-IV. A small number of prototype scores gain disproportional attention, spiking high above the nearby scores. The scores of 50 and 100 stand out again by attracting 5.49% and 5.58% of all score occurrences, respectively. The other five scores exceeding the popularity threshold of 2% are 70, 80, 60, 40, and 90, with them getting 2.67%, 2.66%, 2.57%, 2.42%, and 2.10% of all score occurrences, respectively. The next four scores in order of decreasing popularity are 65, 85, 95, and 75. Similarly to the findings for Waterloo-IV, scores divisible by five emerge as the prototypes of the score categories newly formed by the raters.

Our analyses indicate that 100 levels are clearly excessive for subjective assessment of QoE in ABR streaming, at least by the factor of five given the popularity of scores divisible by five. Raters' responses in the iQoE post-assessment survey support this sentiment. Hence, we align our recommendation on the scoring scale with the perspective in [31], [32] that a small number of levels, e.g., five in the ACR scale, are sufficient for efficient accurate characterization of QoE:

(Test conducting) Use a scoring scale with a small number of levels, such as the five-level ACR scale.

V. Interface Design for Subjective Assessments

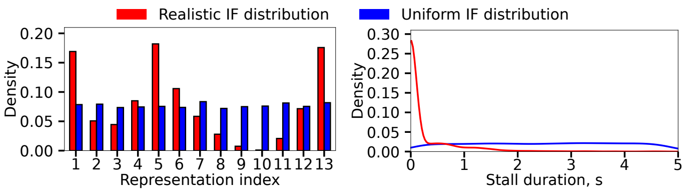
The prominence gained by score 50 in the Waterloo-IV and iQoE datasets deserves a separate discussion. 50 is by far the most popular score in comparison to all other intermediate scores on the 1-100 scale in Figure 1. In the iQoE dataset, this outcome might arise partly due to score 50 constituting, as Figure 14b in [18] shows, the initial position of the handle on the slider in each iQoE assessment. Because keeping the handle in the initial position before submitting the score of 50 is effortless, and the effort to change the score by dragging the handle from 50 to 51, 60, or 61 is about the same and not negligible, it seems logical that the popularity difference between scores 50 and 51 compared to scores 60 and 61 is significantly larger. Although another likely contributor to the dominance of intermediate score 50 is its central role on the 1-100 scale as the middle point in the ternary QoE perspective between the extreme scores of 1 and 100, our previous observation highlights the importance of designing an unbiased interface for subjective tests, e.g., by randomizing the initial position of the handle on the slider in different assessments:

(Test conducting) Design an unbiased interface for subjective assessments, e.g., a randomized initial position of the slider handle.

VI. Experience Selection for Subjective Tests

The outcome of subjective tests depends significantly on the experiences presented to the raters and, in particular, on the IF values of these experiences. Hence, the choice of the tested experiences and their IF values is an important task. However, the following three circumstances complicate the task. First, multiple IFs characterize an experience. Second, an IF might have many potential values, e.g., stall duration spread between 0 and 5 s. Third, a rater is capable of evaluating only a relatively short series of experiences.

Despite the importance, the selection of experiences routinely lacks in sufficient care. Specifically, it is common to select values for an IF across the experience series in a simplistic manner, such as by drawing the values uniformly or randomly from the range of the IF's possible values. While choosing the values for stall duration and frequency in the randomly uniform fashion, [30] employs other ad-hoc rules for video quality. [50] and [51] adopt similar approaches for their IFs of video quality and stalling. [52] restricts stalling to either beginning or middle of experiences.



(a) Representation index

(b) Stall duration

Fig. 2: Realistic, as per the iQoE dataset, and uniform selection of IF values for tested experiences.

To examine experience selection, we utilize the iQoE set of the 1,000 experiences generated through simulations on real network traces with three ABR algorithms and a bitrate ladder comprising 13 representations indexed from 1 to 13. Representation 1 has bitrate 235 Kbps and resolution 320×180. The bitrate and resolution in representation 13 are 16,800 Kbps and 3,840×2,160, respectively. Figure 2a shows that representations 1, 5, and 13 are the most frequent, with each of them taking in this realistic experience set a larger share than 15%. Representations 8 through 11 are the least frequent, with their individual shares in the experience set falling below 3%. The plot contrasts this realistic distribution with the randomly uniform sampling of the representation index between 1 and 13. Figure 2a clearly illustrates that the randomly uniform selection of values for the representation index gives unrealistically high attention to unpopular representations 8 through 11 and unrealistically low attention to popular representations 1, 5, and 13.

We change the IF of interest to stall duration and compare its realistic value distribution in the iQoE experience set against the randomly uniform sampling of stall duration between 0 and 5 s. Figure 2b plots kernel density estimates for the two alternatives. In the realistic distribution, stall duration is predominantly below 0.5 s and rarely exceeds 2 s. Thus, the uniform selection of values for stall duration substantially exaggerates the real stalling behavior.

The above analysis illustrates that uniform and other simplistic approaches to experience selection are unrealistic, thereby endangering the validity of conducted subjective tests. Thus, we give the following advice:

(Test conducting) Realistically select experiences for subjective tests and, in particular, with respect to the IF values across the tested experiences.

VII. Validation of QoE Models

Sections IV, V, and VI demonstrate that subjective tests require a substantial amount of thoughtfulness in their setup in order to appropriately collect scores needed for constructing a QoE model. On the other hand, a QoE model in ABR streaming is rarely a goal in itself and instead receives an auxiliary role in the design, operation, or evaluation of ABR algorithms. This might be a reason why various QoE modeling efforts are insufficiently careful. It is not uncommon to propose a QoE model based on abstract considerations without a proper experimental validation.

QoE model B [33] is a prominent example of the validation concern. The model employs a linear approximation function and combines four IFs as a weighted sum with predefined weight values. [33] introduces QoE model B without conducting subjective tests and does not validate its choices of the linear function and specific IFs. A simple experiment only illustrates how three sets of weight values affect the QoE value produced by the model. Because [33] and [43] leverage QoE model B in their respective MPC and Pensieve algorithms, the success of these pioneering QoE-based ABR algorithms heightens attention to this QoE model and inspires numerous attempts to improve it. The improvements by QoE models G [34], R [35], S [36], V [37], and N [38] primarily target the usage of the bitrate as a proxy of video quality in QoE model B. For example, instead of the bitrate, PSNR and VMAF characterize video quality in QoE models R and V, respectively. However, the above extensions of QoE model B neither question nor validate its major underlying assumptions, such as the linearity of its approximation function.

The existence of the prominent family of QoE models that lack a proper validation leads us to dual recommendations which are both obvious and unfortunately relevant:

(Model building) When proposing a QoE model, validate it through subjective tests.

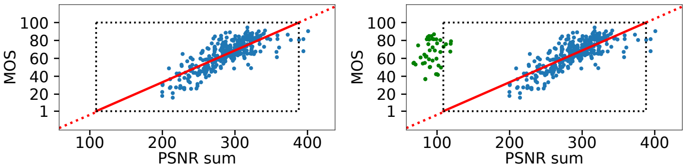
(Model using) Use validated QoE models only.

VIII. Value Interpretability of QoE Models

Lacking validation of a QoE model might have another negative side effect of the model values losing their interpretability. Due to the separation from subjective tests and their scoring scale, the QoE model is likely to yield values that defy interpretation by humans. For example, while Figures 8 through 15 in [43] evaluate different ABR algorithms via three variants of QoE model B, the values produced by the QoE_lin and QoE_hd variants range from -0.5 to 3 and from -1 to 15, respectively, and it remains unclear how these two empirical value ranges relate to the bad, poor, fair, good, and excellent levels of the common scoring scales.

The disconnection of QoE values from their interpretation also undermines their utility for comparison of different ABR algorithms. Although higher values produced by a QoE model typically indicate better quality of experience, the lacking interpretability of QoE values translates into lacking interpretability of their differences, e.g., of whether a QoE increase with a new ABR algorithm is not meaningful due to falling within the just-noticeable difference (JND), i.e., the maximum difference imperceptible by a human [53].

The above problematic example of QoE model B and its variants that produce both negative and positive values calls for a word of caution about reporting only relative changes in QoE. Consider an ABR algorithm achieving a positive QoE value which lies arbitrarily close to zero. If another ABR algorithm surpasses this QoE value by a small amount, the relative QoE increase might be 100%, 1,000%, or higher even when the absolute increase is within the JND, i.e., meaninglessly small.



(a) Construction on Waterloo-IV

(b) Usage on iQoE data

Fig. 3: A regression-based QoE model: (a) values beyond the scale and (b) mismatch between usage and construction.

Our discussion in this section highlights dangers of segregating a constructed QoE model from subjective, humanly interpretable perception of QoE. Hence, we argue for QoE models that support interpretation of their values. Apart from the advantages for QoE evaluation of ABR algorithms, the value interpretability equips QoE models with other strengths, e.g., their direct applicability as synthetic raters [18]. Besides, we contend that the values produced by the QoE model should be positive numbers so as to facilitate their mathematical treatment, including meaningful relative comparisons. While both desired properties hold for the range of QoE values aligned with the five-level ACR scale, we advise the following:

(Model building) Construct a QoE model producing positive interpretable values, e.g., in the range consistent with the five-level ACR scale.

IX. Capping of the Value Range

Whereas Sections VII and VIII expose general problems in the construction of QoE models, we now examine specific technical reasons why these problems arise. Even when a QoE model aspires to align its value range with a humanly interpretable scale, the common reliance on unconstrained regression does not assure such alignment, including on the data used to train the regression. We consider the 450 experiences assessed by the 32 HDTV raters in Waterloo-IV and retain only the experiences devoid of stalling. For ease of exposition, we characterize each of the remaining 326 experiences with a PSNR sum, a new single IF calculated as the sum of the PSNR values across all seven chunks in the experience.

Figure 3a presents a scatter plot of the PSNR sum and MOS for the 326 experiences as blue dots. The graph also depicts as a red line a QoE model constructed on this data via linear regression with the least squares fitting. The solid portion of the line represents the QoE values between 1 and 100, i.e., within the 1-100 scoring scale of Waterloo-IV. The respective range of the PSNR sum is from 108 to 388. However, four experiences in the training data have a larger PSNR sum than 388, and the QoE model returns 100.1, 102.3, 102.3, and 104.6 as the QoE values for these four experiences. Thus, due to the reliance on the unconstrained regression, the constructed QoE model produces values beyond the targeted scale even on the training dataset.

A simple way for a regression-based QoE model to address the problem of unconstrained regression is to cap the regression output to an intended range of values. For example, [6], [35], and [54] apply capping to prevent negative values, with [6] and QoE model R [35] imposing the nonnegative limit on the outputs of Petrangeli model [55] and QoE model B [33], respectively. [56] and [57] align QoE values with the five-level ACR scale by restraining the values to the range from 1.05 to 4.9 for QoE model P and various models from [40], [52], [58], [59], respectively.

An alternative to the output capping is to use an approximation function with built-in adherence to the intended value range. For instance, QoE model L [42] utilizes a hyperbolic tangent function to guarantee values within the range between -1 and 1 and then linearly transforms the guaranteed range to match the five-level ACR scale. [39] configures the exponential function of QoE model F so that the regression always yields values between 1 and 5. [18] creates synthetic raters by adopting a sigmoid function that assuredly produces values between 1 and 100. [60] guarantees QoE values between 0 and 100 by using a sigmoid function as well.

While not advocating a specific method for ensuring that a QoE model produces values within the targeted range, we view such assurances as important for the interpretability of the QoE model and make the following recommendation:

(Model building) Construct a QoE model that assuredly returns values in the intended interpretable range.

X. Mismatch between Usage and Construction

Restricting the values produced by a QoE model to an intended range does not ensure their meaningful interpretation. Figure 3b enhances Figure 3a by adding a scatter plot of the PSNR sum and MOS for the 43 stalling-free iQoE experiences as green dots, where the PSNR sum again refers to the sum of the PSNR values across all chunks in the experience. However, unlike Waterloo-IV with its seven-chunk experiences, the iQoE dataset composes its experiences from four chunks, and the PSNR sum across the 43 stalling-free iQoE experiences varies from 67 to 119. Consequently, the linear QoE model trained on the Waterloo-IV data, i.e., the red line in Figure 3b, returns values within the intended 1-100 range for only four of the 43 experiences. These QoE values are 1.7, 3.8, 4.0, and 4.1, clustering at the bottom of the range. The other 39 experiences receive QoE values smaller than 1 and as low as -14.4. Even with the regression output capped from below by 1, the QoE model characterizes the 43 experiences with values between 1 and 4.1, which is meaninglessly low because the iQoE dataset carefully assembles experiences to cover the entire QoE spectrum from bad to excellent levels.

The observed problem occurs due to the different settings during the usage and construction of the QoE model. While the training Waterloo-IV data contains seven-chunk experiences with the PSNR sum ranging from 200 to 401, the testing iQoE data employs four-chunk experiences with the PSNR sum varying from 67 to 119. Unfortunately, the mismatch between the usage and construction settings is not uncommon. For instance, [6] and [61] use QoE model P in settings that differ from those explicitly assumed in its construction, such as experiences that last less than a minute or contain more than five stalling events.

The mismatch problem becomes graver because many QoE models do not describe their construction settings fully, clearly, or at all. For example, QoE models P [41] and L [42] do not publicly release their training modules. Furthermore, [42] trains QoE model L on three datasets and only vaguely describes the roles played by two of them in the training. Similarly, [55] leaves the construction settings of its Petrangeli model unclear by simply referring to [54] and [62]. Thus, even an entity willing to use QoE models appropriately might be unable to do so because the models do not disclose their construction settings.

We suggest addressing the problem by quenching its fundamental source, i.e., by using a QoE model in settings covered during its construction. Although there are alternative heuristics, such as extrapolation or normalization of IF values, these heuristics rely on simplifying assumptions and have ad hoc applicability. The following dual advice promotes the general fundamental solution:

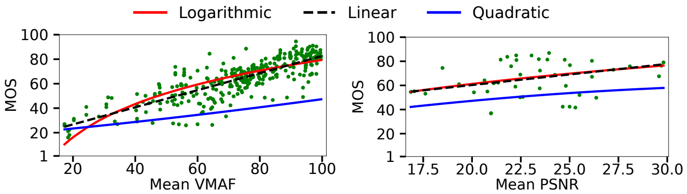
(Model building) Annotate the proposed QoE model with its construction settings.

(Model using) Restrict the usage of QoE models to their annotated construction settings.

XI. Correlation vs. Error

Evaluation of QoE models commonly utilizes metrics of correlation or error. Both kinds of metrics characterize the relationship between the ground-truth subjective scores and values produced by a QoE model. Quantifying the strength and direction of the relationship between these two variables, the correlation metrics include Pearson linear correlation coefficient (PLCC) and Spearman rank correlation coefficient (SRCC), which deal with the two variables' values and their ranks, respectively. Both PLCC and SRCC vary from -1 (perfect negative relationship) through 0 (no relationship) to 1 (perfect positive relationship). On the other hand, mean absolute error (MAE) and root-mean-square error (RMSE) are metrics of error in regression problems and measure differences between the subjective scores and values returned by the QoE model. While MAE treats all individual differences equally, RMSE assigns larger weights to larger differences.

Similarly to Section IX, we utilize the 326 stalling-free Waterloo-IV experiences assessed by the 32 HDTV raters. This time, the only IF is the mean VMAF computed as the average of the VMAF values across all seven chunks in the experience. We apply the Nelder-Mead method [63] to build three regression-based QoE models that employ logarithmic, linear, and quadratic approximation functions. Specifically and respectively for the logarithmic, linear, and quadratic QoE models, we aim to minimize MAE, minimize RMSE, and maximize PLCC with (0, 0), (0, 0), and (2, 2, 1) as the initial simplex.



(a) On Waterloo-IV data

(b) On iQoE data

Fig. 4: Three QoE models constructed via logarithmic, linear, and quadratic regressions.

TABLE I: MAE, RMSE, and PLCC performance of the three logarithmic, linear, and quadratic QoE models.

(a) On Waterloo-IV data

	Logarithmic	Linear	Quadr.
MAE	14.8	15.4	26.9
RMSE	20.5	19.5	30.2
PLCC	0.053	0.105	0.115

(b) On iQoE data

	Logarithmic	Linear	Quadr.
MAE	13.4	13.3	17.2
RMSE	15.8	15.9	20.7
PLCC	0.108	0.103	0.112

Figure 4a depicts the three QoE models along with their training Waterloo-IV data. All three models perform identically with respect to SRCC by achieving the same value of 0.184. Table I-a reports the MAE, RMSE, and PLCC performance of the three QoE models and, for each of the metrics, highlights in orange the cell with the best performance. The results reveal that each of the logarithmic, linear, and quadratic QoE models outperforms the other two counterparts in regard to MAE, RMSE, and PLCC by providing the best values of 14.8, 19.5, and 0.115, respectively. On the one hand, it is not surprising that the QoE model achieving the best value for a metric is the model constructed to optimize this metric. On the other hand, it is remarkable that the performance of this QoE model is never the best in regard to the other metrics.

We also conduct a similar analysis for the logarithmic, linear, and quadratic QoE models trained on the 43 stalling-free iQoE experiences from Section X. The only IF is the mean PSNR calculated as the average of the PSNR values across all four chunks in the experience. To build the logarithmic and linear QoE models, we apply the Nelder-Mead method to minimize MAE with (1, 0) as the initial simplex. For the quadratic QoE model, we strive to maximize PLCC with (2, 0, -1) as the initial simplex.

Figure 4b plots the training iQoE data and three regression-based QoE models. Again, the SRCC performance is the same across the QoE models, with all three models delivering the identical value of 0.127. In regard to MAE, RMSE, and PLCC, Table I-b confirms the qualitative conclusion reached above for Waterloo-IV: while each of the QoE model outperforms its counterparts in one metric, the performance of this QoE model is not the best with respect to the other metrics. Our findings manifest that the advantage of a QoE model in regard to one metric might be misleading and that substantiating the overall goodness of the QoE model necessitates its comprehensive evaluation via multiple metrics.

TABLE II: Average QoE performance of ABR algorithms on the Waterloo-IV dataset according to different QoE models.

	B	G	R	S	V	N	F	A	P	L
Pensieve	37.58	59.22	62.95	61.93	54.60	54.60	66.48	58.03	3.13	3.97
MPC	58.91	71.17	66.07	66.56	67.64	67.64	63.81	69.31	3.93	3.00
BBA	58.92	73.92	64.79	64.97	67.71	67.71	66.54	66.92	3.46	4.73
TR	53.62	63.74	67.90	68.74	69.04	69.04	66.17	69.44	3.88	3.32
Increase, %	0.02	3.86	2.77	3.28	1.96	1.96	0.09	0.19	1.29	19.1

Although the above analyses on the Waterloo-IV and iQoE data indicate that error and correlation metrics, including their MAE, RMSE, and PLCC varieties, are important due to quantifying different relevant aspects of QoE models, it is not unusual for evaluations to omit some of the metrics. For example, [37], [59], and [64] consider correlation metrics only. While [30], [57], and [65] ignore MAE, [52] excludes RMSE. The evaluation in [42] employs only PLCC and RMSE.

On the question which metrics to use, we call for diversity of perspectives. In spite of existing arguments that error metrics are superior to correlation metrics in their utility for evaluation and understanding of QoE models [29], our position is that metrics of both types are pertinent because of their potential to unveil dissimilar conclusions. For the same reason, we advocate using multiple metrics of the same type, e.g., both MAE and RMSE as error metrics. Hence, our recommendation on metrics is as follows:

(Model building) For diversity of perspectives, evaluate QoE models via metrics of both error and correlation, including MAE, RMSE, and PLCC.

XII. QoE Evaluation of ABR Algorithms

Moving the evaluation focus from QoE models to QoE achieved by ABR algorithms, we start by analyzing the usage of QoE models for ABR evaluation. We use 945 (i.e., 70%) of all 1,350 experiences in the Waterloo-IV dataset to train QoE models B, G, R, S, V, N, F, and A, i.e., eight parameterized models from Section II. After the training, these QoE models produce values predominantly within the 1-100 range. The training dataset of 945 experiences includes 189 (i.e., 70%) of the 270 experiences generated with each of Waterloo-IV's five ABR algorithms. To test the achieved QoE, we consider Pensieve, MPC, BBA, and TR as four well-known schemes among these five ABR algorithms. For each of the four tested ABR schemes, we evaluate the average QoE over the remaining 81 (i.e., 30%) of the 270 experiences generated with this ABR scheme. We conduct this QoE evaluation by separately using each of the 10 QoE models from Section II, including models P and L which return values in the 1-5 range. Because QoE models P and L come without public training modules, our evaluation uses these two models in their publicly released configurations without any retraining.

Table II reports the QoE performance achieved by the four ABR algorithms according to the 10 QoE models. For each QoE model, the table highlights in orange and blue the cell with the best and second-best QoE value, respectively, and shows the relative improvement of the former over the latter in the bottom cell. Table II shows that TR provides the highest average QoE according to five of the 10 QoE models, with the relative QoE improvement over the second-best ABR algorithm ranging from 0.19% to 3.28%. BBA delivers the highest QoE according to four QoE models. MPC provides the best average QoE according to QoE model P only. Pensieve is never on top and ends up being the second-best ABR algorithm according to two of the 10 QoE models. The findings show that the choice of a QoE model for evaluation of ABR algorithms significantly affects which of the algorithms achieves the highest QoE. [66] tunes various ABR algorithms for four QoE models and reaches the same conclusion that the ability of an ABR algorithm to outperform its counterparts depends on the QoE model selected for QoE evaluation.

Nevertheless, a widespread practice is to evaluate QoE performance of ABR algorithms by using only one QoE model or a small set of similar QoE models. For example, [43] evaluates QoE under Pensieve vs. other ABR algorithms via three similar variants of QoE model B. The QoE evaluation of STALLION [67] employs a version of QoE model B that accounts for latency. [68] compares QoE of its Stick proposal and baseline ABR algorithms by utilizing differently parameterized instances of a single QoE model.

The concerns about using only one QoE model to evaluate QoE performance of an ABR algorithm get exacerbated when the design or operation of the evaluated ABR algorithm relies on the very same QoE model. Instead of detecting any systematic error introduced into the ABR algorithm by the QoE model, such QoE evaluation espouses and exonerates the bias of this QoE model. Besides, the evaluation gives the ABR algorithm an unfair advantage in comparison with other ABR algorithms that do not employ this QoE model in their design and operation. This bias problem afflicts the evaluations in [33], [37], and [69].

Given the diversity of existing QoE models and the lack of a single, universally accepted QoE model, we argue that QoE evaluation of ABR algorithms should use multiple diverse QoE models. The alternative perspectives offered by multiple QoE models mitigate the biases of individual models and promote comprehensive evaluation of QoE achieved by ABR algorithms. Hence, our recommendation on the usage of QoE models for evaluation of ABR algorithms is as follows:

(Model using) Evaluate ABR algorithms via multiple diverse QoE models.

The shift from QoS to QoE aspires to provide, among other goals, a holistic metric of the user's overall satisfaction. The lack of consensus on the most appropriate QoE model indicates that this aspiration still falls short of its fulfillment. [33], [43], and [70] evaluate QoE performance of ABR algorithms by not only using QoE models but also assessing individual IFs employed by the QoE models. Furthermore, [34], [36], and [71] relinquish QoE models altogether and appraise QoE in the QoS style by evaluating only individual IFs. In this regard, we again follow the spirit of comprehensive evaluation through diversity of perspectives and advise that QoE models and individual IFs meaningfully complement each other in QoE evaluation:

(Model using) To evaluate QoE provided by ABR algorithms, complement usage of QoE models with appraisal of individual IFs.

The constellation of problems that plague objective QoE evaluation of ABR algorithms brings usage of subjective tests back into the spotlight. Despite the larger overhead, direct assessment of QoE via subjective tests is attractive due to its higher accuracy. Although conducting large-scale subjective assessments is not always feasible, we strongly recommend considering this option for QoE evaluation of ABR algorithms:

(Test conducting) Use subjective tests to evaluate QoE achieved by ABR algorithms.

XIII. Conclusions

This paper reviewed the current landscape of QoE in ABR video streaming. Based on two large real datasets of QoE perception by individual raters, we identified and examined various QoE-related pitfalls in test conducting, model building, and model using. Our analyses also derived the following guidelines for improving the status quo:

- **Test conducting:** We recommended scoring scales with a small number of levels (such as the five-level ACR scale), unbiased interface design (e.g., with a randomized initial position of the slider handle), realistic selection of IF values across the tested experiences, and usage of subjective tests to not only build QoE models but also evaluate QoE performance of ABR algorithms.
- **Model building:** Our paper argued that a proposed QoE model should be validated via subjective tests and annotated with its construction settings, that the QoE model should produce positive interpretable values in the intended range, and that evaluation of the QoE model should utilize metrics of both error and correlation (such as MAE, RMSE, and PLCC).
- **Model using:** We suggested usage of validated QoE models and only in their annotated construction settings, as well as evaluation of ABR algorithms via multiple diverse QoE models and individual IFs.

The chief aspiration of this paper was to improve awareness of various problems in the current treatment of QoE and to indicate a way forward. We hope that our observations will help to foster high standards in future work on QoE in ABR video streaming.

Acknowledgments

The following projects support this research:

- TED2021-131264B-I00 (SocialProbing) funded by:
  - MCIN/AEI/10.13039/501100011033 and
  - European Union NextGenerationEU/PRTR,
- PID2021-128223OA-I00 (GreenEdge) funded by:
  - MCIN/AEI/10.13039/501100011033 and
  - European Union ERDF “A way of making Europe”.

References

[1] K. Brunnström et al., “Definitions of Quality of Experience,” *Qualinet White Paper*, 2013.

[2] International Telecommunication Union, “Vocabulary for Performance, Quality of Service and Quality of Experience,” 2017, recommendation P.10/G.100.

[3] A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann, “A Survey on Bitrate Adaptation Schemes for Streaming Media over HTTP,” *IEEE Communications Surveys and Tutorials*, vol. 21, no. 1, pp. 562–585, 2019.

[4] Conviva, “Conviva’s State of Streaming Q2 2022,” September 2022, Report, https://www.conviva.com/wp-content/uploads/2022/09/Q2-SoS.pdf.

[5] Sandvine, “The Global Internet Phenomena Report January 2023,” January 2023, Report, https://www.sandvine.com/global-internet-phenomena-report-2023.

[6] A. Seufert, F. Wamser, D. Yarish, H. Macdonald, and T. Hoßfeld, “QoE Models in the Wild: Comparing Video QoE Models Using a Crowdsourced Data Set,” in *QoMEX 2021*.

[7] F. Chen, C. Zhang, F. Wang, and J. Liu, “Crowdsourced Live Streaming Over the Cloud,” in *INFOCOM 2015*.

[8] U. Reiter et al., “Factors Influencing Quality of Experience,” in *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer, 2014.

[9] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, “A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 2s, p. 1–29, 2018.

[10] T. Zhao, Q. Liu, and C. W. Chen, “QoE in Video Transmission: A User Experience-Driven Strategy,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 285–302, 2017.

[11] P. Juluri, V. Tamarapalli, and D. Medhi, “Measurement of Quality of Experience of Video-on-Demand Services: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 401–418, 2016.

[12] N. Barman and M. G. Martini, “QoE Modeling for HTTP Adaptive Video Streaming – A Survey and Open Challenges,” *IEEE Access*, vol. 7, pp. 30 831–30 859, 2019.

[13] A. A. Barakabitze, N. Barman, A. Ahmad, S. Zadtootaghaj, L. Sun, M. G. Martini, and L. Atzori, “QoE Management of Multimedia Streaming Services in Future Networks: A Tutorial and Survey,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 526–565, 2020.

[14] P. Davis, C. D. Creusere, and J. Kroger, “EEG and the Human Perception of Video Quality: Impact of Channel Selection on Discrimination,” in *GlobalSIP 2013*.

[15] Q. Huynh-Thu and M. Ghanbari, “Scope of Validity of PSNR in Image/Video Quality Assessment,” *Electronics Letters*, vol. 44, no. 13, p. 800–801, 2008.

[16] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward A Practical Perceptual Video Quality Metric,” 2016, Netflix Technology Blog. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[17] F. Tashtarian, A. Bentaleb, H. Amirpour, S. Gorinsky, J. Jiang, H. Hellwagner, and C. Timmerer, “ARTEMIS: Adaptive Bitrate Ladder Optimization for Live Video Streaming,” in *NSDI 2024*.

[18] L. Peroni, S. Gorinsky, F. Tashtarian, and C. Timmerer, “Empowerment of Atypical Viewers via Low-Effort Personalized Modeling of Video Streaming Quality,” in *CoNEXT 2023*.

[19] X. Zhang, Y. Ou, S. Sen, and J. Jiang, “Sensei: Aligning Video Streaming Quality with Dynamic User Sensitivity,” in *NSDI 2021*.

[20] X. Zuo, J. Yang, M. Wang, and Y. Cui, “Adaptive Bitrate with User-Level QoE Preference for Video Streaming,” in *INFOCOM 2022*.

[21] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz, “Quality of Service Terminology in IP Networks,” *IEEE Communications Magazine*, vol. 41, no. 3, pp. 153–159, 2003.

[22] V. Nathan, V. Sivaraman, R. Addanki, M. Khani, P. Goyal, and M. Alizadeh, “End-to-End Transport for Video QoE Fairness,” in *SIGCOMM 2019*.

[23] Z. Duanmu, A. Rehman, and Z. Wang, “A Quality-of-Experience Database for Adaptive Video Streaming,” *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 474–487, 2018.

[24] Z. Duanmu, W. Liu, Z. Li, D. Chen, Z. Wang, Y. Wang, and W. Gao, “Assessing the Quality-of-Experience of Adaptive Bitrate Video Streaming,” *arXiv*, no. 2008.08804, 2020.

[25] D. Z. Rodríguez, R. L. Rosa, and G. Bressan, “Video Quality Assessment in Video Streaming Services Considering User Preference for Video Content,” in *ICCE 2014*.

[26] D. Ghadiyaram, J. Pan, and A. C. Bovik, “A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 183–197, 2019.

[27] International Telecommunication Union, “Methodologies for the Subjective Assessment of the Quality of Television Images,” 2023, recommendation BT.500-15.

[28] ——, “Subjective Video Quality Assessment Methods for Multimedia Applications,” 2022, recommendation P.910.

[29] H. Sheikh, M. Sabir, and A. Bovik, “A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[30] N. Eswara, K. Manasa, A. Kommineni, S. Chakraborty, H. P. Sethuram, K. Kuchi, A. Kumar, and S. S. Channappayya, “A Continuous QoE Evaluation Framework for Video Streaming Over HTTP,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3236–3250, 2018.

[31] 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/ACM Transactions on Networking*, vol. 57, no. 1, pp. 1–14, 2011.

[32] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, “Performance Comparisons of Subjective Quality Assessment Methods for Mobile Video,” in *QoMEX 2010*.

[33] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, “A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP,” in *SIGCOMM 2015*.

[34] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, “BOLA: Near-Optimal Bitrate Adaptation for Online Videos,” in *INFOCOM 2016*.

[35] I. de Fez, R. Belda, and J. C. Guerri, “New Objective QoE Models for Evaluating ABR Algorithms in DASH,” *Computer Communications*, vol. 158, pp. 126–140, 2020.

[36] F. Y. Yan, H. Ayers, C. Zhu, S. Fouladi, J. Hong, K. Zhang, P. Levis, and K. Winstein, “Learning in Situ: A Randomized Experiment in Video Streaming,” in *NSDI 2020*.

[37] T. Huang, C. Zhou, X. Yao, R. X. Zhang, C. Wu, B. Yu, and L. Sun, “Quality-Aware Neural Adaptive Video Streaming with Lifelong Imitation Learning,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2324–2342, 2020.

[38] A. Bentaleb, A. C. Begen, and R. Zimmermann, “SDNDASH: Improving QoE of HTTP Adaptive Streaming Using Software Defined Networking,” in *MM 2016*.

[39] T. Hossfeld, R. Schatz, E. Biersack, and L. Plissonneau, “Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience,” in *Data Traffic Monitoring and Analysis*. Springer, 2013.

[40] C. G. Bampis and A. C. Bovik, “Learning to Predict Streaming Video QoE: Distortions, Rebuffering and Memory,” *arXiv*, no. 1703.00633, 2017.

[41] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, “A Bitstream-Based, Scalable Video-Quality Model for HTTP Adaptive Streaming: ITU-T P.1203.1,” in *QoMEX 2017*.

[42] H. T. T. Tran, D. V. Nguyen, N. P. Ngoc, and T. C. Thang, “Overall Quality Prediction for HTTP Adaptive Streaming Using LSTM Network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3212–3226, 2021.

[43] H. Mao, R. Netravali, and M. Alizadeh, “Neural Adaptive Video Streaming with Pensieve,” in *SIGCOMM 2017*.

[44] T. Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, “A Buffer-Based Approach to Rate Adaptation: Evidence From a Large Video Streaming Service,” in *SIGCOMM 2014*.

[45] K. Spiteri, R. K. Sitaraman, and D. Sparacio, “From Theory to Practice: Improving Bitrate Adaptation in the DASH Reference Player,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2s, p. 1–29, 2019.

[46] Z. Duanmu, W. Liu, Z. Li, D. Chen, Z. Wang, Y. Wang, and W. Gao, “The Waterloo Streaming Quality-of-Experience Database-IV,” IEEE Dataport, 2020, https://dx.doi.org/10.21227/j15a-8r35.

[47] L. Peroni, S. Gorinsky, F. Tashtarian, and C. Timmerer, “iQoE Dataset and Code,” GitHub, 2023, https://github.com/Leo-rojo/iQoE_Dataset_and_Code.

[48] H. Mao et al., “Park: An Open Platform for Learning-Augmented Computer Systems,” in *NeurIPS 2019*.

[49] E. Rosch, “Principles of Categorization,” in *Cognition and Categorization*. Lawrence Erlbaum, 1978.

[50] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, “Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[51] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. C. Bovik, “Modeling the Time – Varying Subjective Quality of HTTP Video Streams With Rate Adaptations,” *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2206–2221, 2014.

[52] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, “A Quality-of-Experience Index for Streaming Video,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 154–166, 2017.

[53] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, “Experimental Design and Analysis of JND Test on Coded Image/Video,” *SPIE Applications of Digital Image Processing XXXVIII*, vol. 9599, pp. 324–334, 2015.

[54] M. Claeys, S. Latré, J. Famaey, T. Wu, W. Van Leekwijck, and F. D. Turck, “Design and Optimisation of a (FA)Q-Learning-Based HTTP Adaptive Streaming Client,” *Connection Science*, vol. 26, no. 1, pp. 25–43, 2014.

[55] S. Petrangeli, J. Famaey, M. Claeys, S. Latré, and F. De Turck, “QoE-Driven Rate Adaptation Heuristic for Fair Adaptive Video Streaming,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 2, pp. 1–24, 2015.

[56] H. Bermúdez-Orozco, J.-M. Martinez-Caro, R. Sanchez-Iborra, J. Arciniegas, and M.-D. Cano, “Live Video-Streaming Evaluation Using the ITU-T P.1203 QoE Model in LTE Networks,” *Computer Networks*, vol. 165, 2019.

[57] D. Nguyen, N. Pham Ngoc, and T. C. Thang, “QoE Models for Adaptive Streaming: A Comprehensive Evaluation,” *Future Internet*, vol. 14, no. 5, 2022.

[58] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, “Deriving and Validating User Experience Model for DASH Video Streaming,” *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 651–665, 2015.

[59] Z. Duanmu, W. Liu, D. Chen, Z. Li, , Z. Wang, Y. Wang, and W. Gao, “A Knowledge-Driven Quality-of-Experience Model for Adaptive Streaming Videos,” *arXiv*, no. 1911.07944, 2019.

[60] A. V. Ivchenko, P. A. Kononyuk, A. V. Dvorkovich, and L. A. Antiufrieva, “Study on the Assessment of the Quality of Experience of Streaming Video,” in *SYNCHROINFO 2020*.

[61] B. Taraghi, M. Nguyen, H. Amirpour, and C. Timmerer, “Intense: In-Depth Studies on Stall Events and Quality Switches and Their Impact on the Quality of Experience in HTTP Adaptive Streaming,” *IEEE Access*, vol. 9, pp. 118 087–118 098, 2021.

[62] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, “Model for Estimating QoE of Video Delivered Using HTTP Adaptive Streaming,” in *IM 2013*.

[63] F. Gao and L. Han, “Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters,” *Computational Optimization and Applications*, vol. 51, pp. 259–277, 2012.

[64] Z. Duanmu, W. Liu, D. Chen, Z. Li, Z. Wang, Y. Wang, and W. Gao, “A Bayesian Quality-of-Experience Model for Adaptive Streaming Videos,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3s, pp. 1–24, 2023.

[65] N. Eswara, S. Ashique, A. Panchbhai, S. Chakraborty, H. P. Sethuram, K. Kuchi, A. Kumar, and S. S. Channappayya, “Streaming Video QoE Modeling and Prediction: A Long Short-Term Memory Approach,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 661–673, 2020.

[66] Y. Liu and J. Y. B. Lee, “A Unified Framework for Automatic Quality-of-Experience Optimization in Mobile Video Streaming,” in *INFOCOM 2016*.

[67] C. Gutterman, B. Fridman, T. Gilliland, Y. Hu, and G. Zussman, “STALLION: Video Adaptation Algorithm for Low-Latency Video Streaming,” in *MMsys 2020*.

[68] T. Huang, C. Zhou, R.-X. Zhang, C. Wu, X. Yao, and L. Sun, “Stick: A Harmonious Fusion of Buffer-Based and Learning-Based Approach for Adaptive Streaming,” in *INFOCOM 2020*.

[69] B. Alt, T. Ballard, R. Steinmetz, H. Koeppl, and A. Rizk, “CBA: Contextual Quality Adaptation for Adaptive Bitrate Video Streaming,” in *INFOCOM 2019*.

[70] Z. Akhtar, S. Rao, B. Ribeiro, Y. S. Nam, J. Chen, J. Zhan, R. Govindan, E. Katz-Bassett, and H. Zhang, “Oboe: Auto-Tuning Video ABR Algorithms to Network Conditions,” in *SIGCOMM 2018*.

[71] C. Wang, A. Rizk, and M. Zink, “SQUAD: A Spectrum-Based Quality Adaptation for Dynamic Adaptive Streaming over HTTP,” in *MMSys 2016*.