New objective QoE models for evaluating ABR algorithms in DASH

Ismael de Fez[a], Román Belda[a], Juan Carlos Guerri[a]

[a] Institute of Telecommunications and Multimedia Applications (iTEAM), Universitat Politècnica de València, Camino de Vera, 46022, Valencia, Spain

e-mail: isdefez@iteam.upv.es, robelor@iteam.upv.es, jcguerri@dcom.upv.es
phone: 34-963879588

Abstract. As users become more demanding with regards to the consumption of multimedia content, the importance of measuring their level of satisfaction is growing. The difficulty in terms of time and resources for assessing the Quality of Experience (QoE) has popularized the use of objective QoE models, which try to emulate human behavior regarding the playback of multimedia streaming. Some objective QoE models existing in the literature are based on the bitrate. However, the PSNR (Peak Signal-to-Noise Ratio) or VMAF (Video Multimethod Assessment Fusion) have been proved to be metrics with a closer relationship with the QoE than the bitrate. This paper proposes three new models to measure the QoE analytically in DASH (Dynamic Adaptive Streaming over HTTP) video services. The first is based on the bitrate of the displayed video segments, whereas the second and the third are based on the PSNR and VMAF of each video segment, respectively. The proposed models are compared to the ITU-T standard P.1203 as well as the bitrate-based QoE model proposed by Yin et al. Moreover, the paper presents a subjective study, which confirms the validity of the proposed models. The models are validated by using different DASH adaptation algorithms. In this sense, this paper also presents a DASH ABR (Adaptive Bitrate Streaming) algorithm called Look Ahead, which takes into account the inherent bitrate variability of the video encoding process in order to calculate, in real time, the appropriate quality level that minimizes the number of stalls during the playback.

Keywords: Quality of Experience (QoE), Dynamic Adaptive Streaming over HTTP (DASH), Peak Signal-to-Noise Ratio (PSNR), Video Multimethod Assessment Fusion (VMAF), Adaptive Bitrate Streaming (ABR), ITU-T P.1203

1. Introduction

We can safely state that video is quite popular on the Internet [1], and sites like YouTube and Netflix accumulate a good share of the Internet traffic [2]. This growth is based on the CDN (Content Delivery Network) friendliness of HTTP (Hypertext Transfer Protocol) adaptive streaming (HAS) [1] techniques like HTTP Live Streaming (HLS) [3] and Dynamic Adaptive Streaming over HTTP (DASH) [4]. Nowadays, HAS has become the most important example of adaptive streaming, and it is used by the aforementioned platforms YouTube and Netflix, among others.

HAS is based on the existence of different qualities of the same audiovisual content, so users consume the content with a certain quality in each moment depending on the client context such as measured bandwidth, device type or screen resolution, among others. During the playback, the displayed quality of the content can change.

Therefore, the content playback may not be the same for two different viewers, so each user may have a different experience, as a consequence of each playback. Hence, the interest of over-the-top media service providers for harvesting data related to the satisfaction of their clients.

Quality of Experience is a subjective evaluation parameter to estimate the overall quality of the service provided from the point of view of the user. The importance of this measure has grown in the last years because of the increasing need for providing a good user experience in many services, especially in video streaming.

QoE comprehends everything that may affect the perceived experience when watching a video. Consequently, everything from video quality to room lighting may affect it. Leaving aside external parameters, QoE can be simplified as a function of the quality of the video and playback events. When consuming HAS content, quality changes and video playback stalls adds an additional level of complexity to Video Quality Assessment (VQA) techniques. In this context, this paper proposes various objective methods to measure the QoE of DASH video playbacks for on-demand video services based on different VQA techniques.

This work is initially based on the QoE model proposed by Yin et al. in [5] which merely rely on the average bitrate of the available qualities and extends it to use: 1) bitrates of individual segments; 2) PSNR of segments; and 3) VMAF of segments. Different ABR algorithms are used to evaluate the performance of the proposed QoE models. One of the contributions of the paper is an ABR algorithm called Look Ahead, which takes into account the bitrate of forthcoming segments when choosing the next video representation in order to avoid stalls during the video playback.

The rest of the paper is organized as follows: Section 1 presents the introduction, contributions and limitations. Section 2 presents the related state of the art. Section 3 details the QoE models proposed. Section 4 presents the proposed Look Ahead ABR algorithm for DASH. Section 5 explains the methodology used to carry out the evaluations presented in Section 6. Finally, Section 7 details the conclusions and future work.

1.1 Contributions

The main contributions of this paper are:

- The proposal of three new objective QoE models, one of them based on bitrates of individual segments, another based on PSNR and the other based on VMAF. All three are simple QoE models that consider the main parameters that affect QoE (encoding quality, rebufferings, and quality switchings). Also, the PSNR-based and the VMAF-based QoE models consider the initial loading delay.
- The proposal of an ABR algorithm for DASH called Look Ahead, which is proved to reduce the number and duration of stalls, providing good values of QoE.

Apart from these contributions, it is worth highlighting the main strengths of this work:

- The proposed QoE models are straightforward for users to estimate the QoE of a video playback easily, thanks to their simplicity and to the availability of a program developed by the authors (available in GitHub [6]) that encodes videos and calculates the bitrate, the PSNR and the VMAF of each segment for each representation. As stated in [7], many papers that propose new QoE models lack specific information that limits their reproducibility and comparability.

- The execution of a subjective evaluation to prove the validity of the proposed QoE models, as well as the comparison with well-known models such as the ITU-T P.1203 recommendation [8].
- The use of larger videos to perform the subjective evaluation. Usually, subjective studies are carried out using videos of short duration (less than 1 minute) [7], where it is difficult to estimate how parameters like rebuffering or quality switching affect the QoE. This paper uses videos with a duration larger than 10 minutes.
- The use of VP9 to perform the tests, the latest codec developed by Google and used nowadays in platforms like YouTube, instead of the classic H.264, which is used by most of the works that propose QoE models.
- The evaluation of the proposed algorithm using a real implementation instead of simulations. To that extent, Look Ahead has been implemented and integrated into ExoPlayer v2, the latest version of the library developed by Google to play DASH content on the Android platform. On mobile devices, content providers usually prefer to deliver their content throughout a native app rather than a web app. On the Android platform, that means that providers like Netflix, HBO, Youtube, and many others use ExoPlayer2 as a base player.

1.2 Limitation

In order to check the performance of the QoE models proposed in this paper, we use different DASH adaptive bitrate algorithms. In this sense, one of the main difficulties of this work has been to choose the ABR algorithms to carry out the evaluation, which leads to one of the main limitations of this work. In the literature there are hardly works which provide enough detail to implement the proposed algorithms. One of the works with a public implementation is "dash.js," which is used as the basis for many ABR algorithms. However, in the scenario proposed in this paper, this implementation and all the solutions based on “dash.js” are not directly portable to ExoPlayer (the library developed by Google to play DASH contents and used in this work), since “dash.js” is only browser-oriented and not compatible with ExoPlayer.

Moreover, in order not to add an additional complexity to the studies presented by evaluating more parameters, in the videos used to perform the evaluation we have fixed the resolution and the segment size. Using a fixed resolution (in this case, 1080p) accomplishes the primary purpose of having the same quality/bitrate curve for all algorithms, and it also simplifies the evaluation of the QoE models proposed, which would require to upsample lower resolutions. On the other hand, we have fixed segment size to 10 seconds, although for the future work we have planned to evaluate other segment lengths. In this regard, the conclusion section contains some possible improvements to the proposed QoE models as part of the future work.

2. State of the art

There are two different families of tests for video quality assessment, namely subjective tests –using test subjects to obtain QoE evaluation of video sequences– and objective tests –using algorithms that estimate the quality of the video–. The following sections introduce both types.

2.1 Subjective tests

The International Telecommunication Union (ITU) has published different recommendations that provide a methodology to conduct subjective evaluations formally. For instance, the ITU-R BT.500 recommendation [9] describes several methods to standardize subjective tests, containing procedures and requirements to choose and configure adequate displays, select test subjects, or determining optimum test and reference video sequences. In the same line, a more modern recommendation is the ITU-T P.913 [10], which is an evolution of ITU-T P.910 [11] and ITU-T P.911 [12]. This recommendation describes non-interactive subjective assessment methods for evaluating the audiovisual quality for applications such as Internet video and distribution quality video.

Both recommendations provide different methodologies to calculate the QoE of users. In this regard, one of the most used techniques to measure QoE is the Mean Opinion Score (MOS), in which different users value their experience with regards to a video playback analyzing specific parameters by using a scale between 1 (lowest satisfaction) and 5 (highest satisfaction). The MOS is then generated as the average over a set of subjective evaluations provided by the test audience.

There are different ways of calculating the MOS; the simplest methodology is known as Single-Stimulus (SS). When using this method, the test population provides their score based on a single visualization of the content. To this category belong different strategies as ACR (Absolute Category Rating), SSCQE (Single Stimulus Continuous Quality Rating), SAMVIQ (Subjective Assessment of Multimedia Video Quality) o MUSHRA (Multi-Stimuli with Hidden Reference and Anchor Points).

Apart from the MOS, another important measure is the DMOS (Differential MOS), in which a stimulus is compared to a reference stimulus. Among the strategies used to evaluate the DMOS, we find the Double-Stimulus Continuous Quality Scale (DSCQS), DCR (Degradation Category Rating) or CCR (Comparison Category Rating).

2.2 Objective tests

In general, the main drawback of subjective tests is the time and resources (in terms of number of people) required to carry out the measurements. This motivates the existence of objective tests, which are performed by algorithms that estimate what the opinion of users would be if they were asked for.

The encoding process, the initial loading delay, the ability of HAS to change the quality for each segment, and the inevitable possibility of running out of buffer during the playback are key QoE estimators for evaluating the quality of this kind of services [13].

The literature repeatedly uses these factors to formulate different QoE methods (for example, [5]), although there are works that consider other parameters, such as [14], that studies the impairments related to the frequency and duration of the stalls. These works have a common point in defining the QoE as a formula where the impairments referred to initial delay, playback stalls, and quality changes penalize QoE.

Although [14] states that the number of stalls is also important to determine the QoE, in the present work we take into consideration the total stall time as it has a direct relation with playback stall impairment and it is simpler to compare between different Video Quality Assessment (VQA) techniques.

In this sense, there are three main categories of VQA techniques [15]: no-reference (NR), reduced-reference (RR) and full-reference (FR), depending on if they do not use any reference, if they use partial information of the reference or if the full reference video is used for the quality assessment, respectively. VQA can also be classified based on the context it can be used [16][17], as out-of-service and in-service. The VQA is out-of-service if there is no time constraint, and the reference video is available, whereas in in-service there are strict time constraints.

In the literature, we can find different QoE models of the categories mentioned above. As an example of a prediction model based on a RR VQA, [18] proposes a machine-learning-based QoE estimator that uses STRRED [19] as VQA. In the ITU-T P.1203 recommendation [8], used in this work for model comparison, in order to estimate the QoE it is used an NR VQA that considers different bitstream data and metadata, depending on the mode. Several works existing in the literature use the ITU-T P.1203 recommendation, such as [20] and [21].

Likewise, some works propose QoE models based on the bitrate, such as [5], and others based on the PSNR [22]. Specifically, [22] proposes a linear model based on differential PSNR. Unlike [22], the PSNR-based QoE model proposed in this paper takes into account video stall events, which have been proved to be a relevant parameter regarding the QoE of users.

Finally, a complete and recent study of QoE modeling for HTTP adaptive video streaming can be found in [7], which surveys the key QoE models for HAS applications. The paper identifies and classifies some of the most relevant QoE models existing in the literature in four categories: parametric models, media-layer models, signal-based models, and hybrid models. Regarding the first, parametric models use measured packet/network-related parameters to estimate the quality, mainly rebuffering duration, bitrate, quality switches, or initial loading delay. In this category, we highlight the model proposed by Rodríguez et al. [23], which takes into account the number, length, and location of the rebuffering events, among other parameters. However, the most relevant models belong to the category of hybrid models, which use much more information as input compared to other models (such as bitstream models or packet or network parameters). In this category, we highlight the following models: Liu et al. [24] propose a no-reference QoE taking into account both spatial and temporal quality considering factors such as rebuffering, quality switching, and initial delay; Garcia et al. [25] present a long-term QoE model by using short-term audiovisual quality models (which considers parameters such as GoP, frame rate or rebuffering events); Duanmu et al. [26] propose a QoE prediction approach that accounts for the instantaneous quality degradation due to perceptual video presentation impairment, the playback stalling events, and the instantaneous interactions between them; and finally, it is worth mentioning the works by Bampis and Bovik ([18], [27], [28]), who propose different machine-learning-based models which use objective metrics, rebuffering related factors, and memory-related functions to predict the end-user QoE.

Among the existing QoE models, in this paper, we have chosen the Yin et al. model [5] and the ITU-T P.1203 recommendation [8] to carry out the comparison regarding the proposed QoE models. The reason is that the Yin et al. QoE model is easy to use, direct, highly cited in the literature, takes into account relevant parameters such as stalls and video quality, and also because, in contrast to many papers previously mentioned, [5] provides enough detail to implement the model. On the other hand, the ITU-T P.1023 [8] has been chosen because of its relevance, and since its implementation is feasible thanks to its specification and due to available open-source implementations. In general,

QoE models existing in the literature are not straightforward for users to estimate the QoE of a video playback easily since most are based on complicated formulas. In contrast, the QoE models proposed in this paper are easily usable.

3. Quality of Experience Models

In this section, the different QoE models proposed in this work are presented. First, the Yin et al. QoE model [5] is briefly explained.

3.1 Normalized QoE model

Yin et al. [5] propose a formula where the QoE is calculated through the sum of the QoE of each segment. Thus, Yin et al. define the QoE of video segment 1 through *K* by a weighted sum of three components: video quality, quality variations, and total rebuffering time, shown in (1).

$$QoE_1^K = \sum_{k=1}^{K} q(R_k) - \lambda \sum_{k=1}^{K-1} |q(R_{k+1}) - q(R_k)| - \mu \sum_{k=1}^{K} \left(\frac{LR_k}{C_k} - B_k\right),\ R_k \in \Re\,, \tag{1}$$

where K is the number of segments of the video, $R_k \in \Re$ (where $\Re$ is the set of all available bitrate levels) is the bandwidth of the selected representation of segment k, $q(\cdot)$ is an increasing function which maps selected bitrate R_k to video quality perceived by user $q(R_k)$, L is the duration (in seconds) of each segment, C_k is the average download speed of segment k, B_k is the buffer occupancy at the instant of time when the segment k is being downloaded, and finally, λ and μ are positive weighting parameters corresponding to video quality variations and rebuffering time, respectively.

Regarding these latest parameters, a small λ implies that the user is not particularly concerned about video quality variability, whereas a large μ indicates that the user is deeply concerned about rebuffering. As stalls, generally, disturb users much more than video quality changes do, the value of μ is usually much higher than λ.

Yin et al. define a normalized QoE model to compare the performance of algorithms to the theoretical optimum, calculated assuming that the future bandwidth is known, as (2) shows:

$$nQoE_1^K = \frac{QoE_1^K}{QoE_{opt}}. \tag{2}$$

3.2 QoE model modified

Initially, this paper proposes a modification regarding the QoE model defined by Yin et al. [5]. The proposed model is shown in (3):

$$QoE_1^K = \sum_{k=1}^{K} q(R_k) - \lambda \sum_{k=1}^{K-1} |q(R_{k+1}) - q(R_k)| - \mu \sum_{k=1}^{K} \left(\frac{LR_k}{C_k} - B_k\right),\ R_k \in \Re_S. \tag{3}$$

Although it seems the same formula proposed in [5], there is a meaningful difference. In (1) $R_k \in \Re$, whereas in the proposed formula $R_k \in \Re_S$, where $\Re_S$ has a different set of values for each segment compared to $\Re$. That is, R_k does not belong to a set of available bitrates specified in the Media Presentation Description (MPD) of the DASH video standard, since the bitrate of each representation, generally, changes for every segment. For example, suppose a video encoded with only one quality, for instance with a bitrate of 500 kbps, the value of $\Re$ will always be 500 kbps for each

segment, whereas $\Re_S$ could have different values in each segment around the average bitrate (e.g., 481, 497, 504 kbps…).

Taking into consideration the specific bitrate of each segment instead of the average bitrate of every representation makes the proposed Yin et al.-based QoE model more accurate. And this is because even for constant bitrate encoded videos, the bitrate changes over time, so every single segment of a representation has, almost inevitably, some variation.

3.3 PSNR-based QoE model

The PSNR is an objective metric with proved correlation with the QoE: as the PSNR increases, the QoE improves [29]. Since the bitrate and the PSNR have an increasing logarithmic relationship, bitrate variations do not affect equally the PSNR depending on the value of the bitrate. For example, a slight bitrate variation can imply a high PSNR variation for low bitrates [30].

Both the bitrate and the PSNR are significant objective measures, but the PSNR offers a more representative relationship regarding the QoE. Also, the nonlinearity of the bitrate regarding the PSNR could make think that the QoE model proposed by Yin et al., based on the bitrate, can be improved. Although the Yin et al. model is indeed based on an increasing *q(·)* function that affects the bitrate, and therefore this could be a linear or logarithmic function (among others), this function is not specified in the proposal by Yin et al. [5].

In this sense, this paper proposes a new QoE model based on such an important parameter as the PSNR is, shown in (4):

$$QoE'_{PSNR} = \frac{1}{K}\sum_{k=1}^{K} PSNR(\xi_k) - \zeta \frac{1}{K-1}\sum_{k=1}^{K-1}|PSNR(\xi_{k+1}) - PSNR(\xi_k)| - \eta \cdot 10\log_{10}\left(1 + \frac{1}{d}\sum_{k=1}^{K}\left[\frac{LR_k}{C_k} - B_k\right]\right) - \delta \cdot 10\log_{10}(1 + T_s),\ R_k \in \Re_S, \quad (4)$$

where K is the number of segments of the video, ξ_k is the selected representation of segment k, $PSNR(\xi_k)$ is the PSNR of the selected representation of segment k, d is the total duration of the video (in seconds), L is the duration (in seconds) of each segment, $R_k \in \Re_S$ is the bandwidth of the selected representation of segment k, C_k is the average download speed of segment k, B_k is the buffer occupancy at the instant of time when the segment k is being downloaded, T_s is the start-up delay, and finally ζ, η and δ are positive weighting parameters corresponding to video representation switches, rebuffering time and start-up delay, respectively. As in the previous case, R_k does not belong to a set of available bitrate levels specified in the MPD.

To establish a lower bound in case there are many stalls (so as to avoid negative values of the model), QoE_{PSNR} is defined as follows:

$$QoE_{PSNR} = \max(QoE'_{PSNR}, 0). \quad (5)$$

The structure and idea of the proposed formula are similar to the QoE model by Yin et al., that is: the PSNR increases the value of the QoE model, whereas both the number of representation switches and the rebuffering duration penalize the QoE. The proposed formula also includes the effect of the start-up delay. Conceptually, (4) can be expressed as shown in equation (6), where the stalling ratio is defined as the amount of time spent so that video playback is stalled (rebuffering time) divided by the total duration of the video:

$$QoE'_{PSNR} = Average\ PSNR - \zeta * Average\ PSNR\ switches - \eta \cdot 10\,log_{10}(1 + stalling\ ratio) - \delta \cdot 10\,log_{10}(1 + start\_up\ delay). \quad (6)$$

Note that, when the rebuffering time is zero, the third term of formula (6) will also be, thus the rebuffering will not penalize the QoE. It is important to highlight that, in contrast to the formula developed by Yin et al., the model proposed in this work has units, specifically dB, as the four elements of the sum are expressed in dB. The maximum value of the QoE_{PSNR} model is the maximum average PSNR of the video. Thus, in the ideal case of a video playout without rebufferings, representation switches, and start-up delay, the value of the QoE_{PSNR} will be the PSNR of the selected representation. On the other side, values of the QoE_{PSNR} near zero, produced mainly by high values of rebuffering duration, indicate a rather poor (and unacceptable) user experience.

Therefore, the value of the proposed QoE_{PSNR} model is bounded: upper bounded by the maximum PSNR of the video (which depends on the video) and lower bounded by 0. Thus, the QoE_{PSNR} model can provide information by itself about the Quality of Experience of the video playback, without the need for comparison with other values, in contrast to the model proposed by Yin et al., which is normalized with an ideal case. It is important to note that, as the PSNR model is content-dependent, the QoE_{PSNR} model will also be.

The main difficulty of using the proposed formula is calculating the PSNR of each segment for each representation. This could imply a considerable processing time, which grows as the number of representations increases. To ease this procedure, the authors of this paper have published a program in GitHub (available in [6]) that encodes videos and calculates the PSNR of each segment for each representation.

In order to check how η affects the QoE_{PSNR}, Fig. 1 shows the QoE_{PSNR} for different values of the percentage of rebuffering time regarding the total duration of the video (that is, the stalling ratio) and different values of η. In the figure, the parameter of average PSNR has been fixed to 44 dB, the average PSNR variation has been set to 4 dB, ζ=1, and δ=0, so, in case of no stalls, the QoE_{PSNR} obtained is 40 dB, as the figure shows. As can be seen, the parameter η has a significant impact on the QoE_{PSNR}. For example, when η=5, if the duration of the stalls is 3% of the video playback, a very poor QoE_{PSNR} (10 dB) is obtained. In contrast, when η=2 the same percentage of stalls duration leads to an acceptable value of QoE_{PSNR}=28 dB.

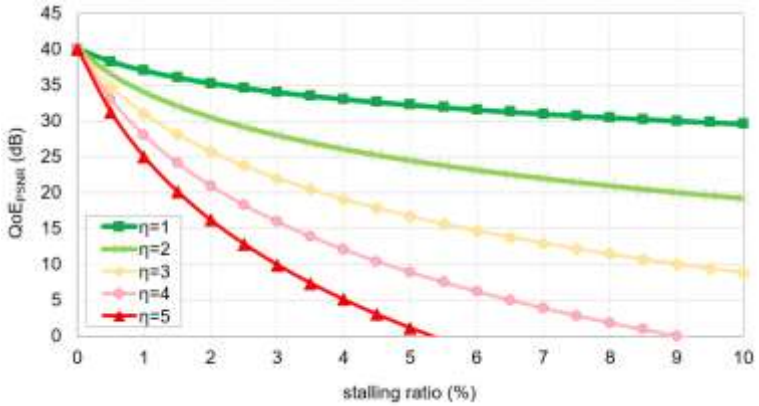


Fig. 1 QoE_{PSNR}(dB) for different values of stalling ratio and η.

3.4 VMAF-based QoE model

VMAF [31][32] is a VQA method developed by Netflix and used by many tools like FFmpeg and Elecard StreamEye. It uses Visual Information Fidelity (VIF) [33], Detail Loss Metric (DLM) [34], and Temporal Impairment Feature (TI) metrics fused by Support Vector Machine (SVM) regression [35] with a built-in machine-learning trained model. The model has been trained using the opinion scores obtained through a subjective experiment, as shown in Fig. 2.

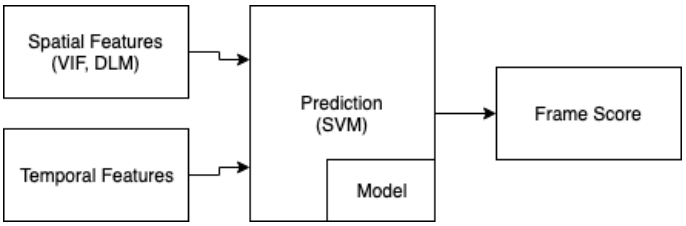


Fig. 2 Outline of the VMAF system.

VMAF adopts a modified version of VIF that uses each one of the values of the four scales used by VIF, while VIF combines them into a single value. The SVR (Support Vector Regression) model uses the six features to generate per-frame value. The final VMAF value is the arithmetic mean of the per-frame values.

The VMAF-based QoE model proposed in this paper is shown in equation (7):

$$QoE'_{VMAF} = \frac{1}{K}\sum_{k=1}^{K} VMAF(\xi_k) - \lambda \frac{1}{K-1}\sum_{k=1}^{K-1} |VMAF(\xi_{k+1}) - VMAF(\xi_k)| - \gamma \cdot \frac{1}{d}\sum_{k=1}^{K}\left[\frac{LR_k}{C_k} - B_k\right] - \delta \cdot T_s,\ R_k \in \Re_S, \quad (7)$$

where $VMAF(\xi_k)$ is the VMAF of the selected representation of segment k, λ and γ are positive weighting parameters corresponding to video quality variations and rebuffering time, respectively, and the rest of parameters are the same as those shown in equation (4).

It is important to highlight that the QoE_{VMAF} model, as the QoE_{PSNR} model, can provide information by itself about the Quality of Experience of the video playback. Thus, the proposed formula has the same scale of VMAF, that is, the maximum value is 100 (an excellent QoE), and the minimum value is 0 (very bad QoE). To that extent, we establish a lower bound defining QoE_{VMAF} as:

$$QoE_{VMAF} = \max(QoE'_{VMAF}, 0). \quad (8)$$

Conceptually, (7) can be expressed as shown in equation (9):

$$QoE'_{VMAF} = Average\ VMAF - \lambda * Average\ VMAF\ switches - \gamma \cdot stalling\ ratio - \delta \cdot start\_up\ delay. \quad (9)$$

Again, in practice, the main difficulty of using the formula is to calculate the VMAF of each segment for each representation. This can be easily calculated by using the program publicly available in GitHub [6], which has been developed by the authors.

To see a similar example to the one shown in Fig. 1, making use of equation (7), Fig. 3 shows the QoE_{VMAF} for different values of stalling ratio and different values of γ. In the figure, the parameter of average VMAF has been fixed to 95, the average VMAF variations have been set to 5, λ=1, and δ=0 (the start-up delay is not considered) so, in case of no stalls, the QoE_{VMAF} obtained is 90. Fig. 1 shows that the parameter γ has a high impact on the QoE_{VMAF} model. To see an example, when γ=1800, if the duration of the stalls is 4% of the video playback, we obtain a very poor

QoE_{VMAF}=18. On the contrary, when γ=600, the same percentage of stalls duration causes an acceptable value of QoE_{VMAF}=66.

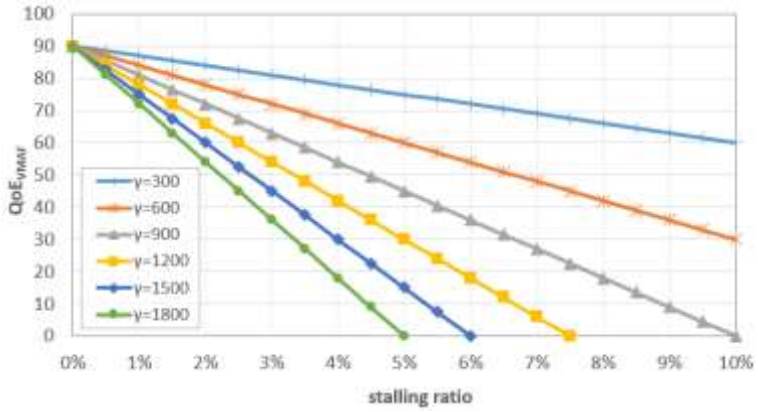


Fig. 3 QoE_{VMAF} for different values of stalling ratio and γ.

4. Look Ahead ABR algorithm

This section proposes an ABR algorithm called Look Ahead that takes into account the bitrate variability of the different qualities and segments. The main objective of Look Ahead is to provide a continuous playback while maximizing video quality. In this way, when calculating the representation chosen for the next segment $i+1$, it is intended to provide the maximum quality, as long as no stalls occur, among the k available representations $Q=\{q_0, q_1, \ldots, q_{k-1}\}$, where q_j is the representation of the segment j (note that $q_j < q_{j+1}$).

Look Ahead is an iterative process that computes the average bandwidth of the forthcoming z segments for all representations from z=1 to θ, where θ is the maximum number of forthcoming segments into consideration to calculate the average rate. On each iteration, the algorithm selects the highest representation of which average bitrate of the next z segments, τ_z, is lower than the estimated bandwidth, according to (10).

$$\tau_z(i+1,j) = \frac{\sum_{m=i+1}^{i+z} S_{m,q_j}}{\sum_{m=i+1}^{i+z} t_m},\ \tau(i+1,j) < \widehat{bw}, \tag{10}$$

where i is the current segment, S_{m,q_j} is the size of segment m for the representation q_j, t_m is the duration of segment m, and $\widehat{bw}$ is the estimated bandwidth. The parameter t_m will usually be equal in every segment, although it depends on the encoding process.

Finally, the lowest representation obtained from the θ iterations is chosen as the selected representation for the next segment, $\xi(i+1)$, shown in (11):

$$\xi(i+1) = min\{\tau_z\},\ z = 1 \ldots \Theta. \tag{11}$$

The fact of considering different iterations when calculating the representation of the next segment is a conservative process that avoids stalls during the playback, since future segments with higher bitrate can make the algorithm to select lower representations than the bandwidth may allow, thus keeping or increasing the buffer depending on forthcoming segments. In this sense, the parameter θ could have a great impact on the QoE of users. In fact, when choosing θ to maximize the QoE, there is a trade-off in terms of stalls, video representation displayed, and noticeable representation switches.

To see an example, in the particular case of θ=3, when calculating the representation of segment *u*, the available rates for the *k* representations under consideration are first calculated. This means calculating $\tau(u,j)$, where $j \in [0, k-1]$, and generating vectors $T(u)$ as shown in (12)-(14):

$$\mathrm{T}(u)_{z=1} = \left[\frac{S_{u,0}}{t_u} \quad \frac{S_{u,1}}{t_u} \quad \ldots \quad \frac{S_{u,k-1}}{t_u}\right], \tag{12}$$

$$\mathrm{T}(u)_{z=2} = \left[\frac{S_{u,0}+S_{u+1,0}}{t_u+t_{u+1}} \quad \frac{S_{u,1}+S_{u+1,1}}{t_u+t_{u+1}} \quad \ldots \quad \frac{S_{u,k-1}+S_{u+1,k-1}}{t_u+t_{u+1}}\right], \tag{13}$$

$$\mathrm{T}(u)_{z=3} = \left[\frac{S_{u,0}+S_{u+1,0}+S_{u+2,0}}{t_u+t_{u+1}+t_{u+2}} \quad \ldots \quad \frac{S_{u,k-1}+S_{u+1,k-1}+S_{u+2,k-1}}{t_u+t_{u+1}+t_{u+2}}\right]. \tag{14}$$

In each vector $T(u)$, the chosen representation is the highest *j* (the highest column in the $T(u)$ vector) that fulfills the condition $\tau(u,j) < \widehat{bw}$, i.e., the necessary rate for downloading that segment must be lower than the estimated bandwidth. When all vectors $T(u)_{z=1..\theta}$ are calculated generating a vector $T_q(u)=\{q_{(z=1)}, q_{(z=2)}, \ldots, q_{(z=\theta)}\}$, the chosen representation of segment *u* will be the lowest representation of the vector $T_q(u)$, i.e. $\xi(u) = min\{T_q(u)\}$.

Note that, when calculating the quality of the last segment, the value of *z* is 1. In the case of the penultimate segment, the value of *z* is $min\{2, \theta\}$, and so on.

5. Methodology

To carry out the evaluation of the different QoE models, among the several ABR algorithms existing in the literature, we have selected, apart from the proposed Look Ahead, Müller [36] and SARA [37] ABR algorithms. The selected algorithms, unlike most papers that propose ABR algorithms for DASH, are reproducible, thus providing enough detail to implement and integrate these algorithms into a real player.

The aforementioned ABR algorithms have been developed and integrated into the ExoPlayer v2 library, the latest version of the library developed by Google to play DASH contents. The use of a real player, instead of emulations, has several advantages for gathering precise data. For example, when using a real implementation, buffer occupancy is updated as soon as a frame is parsed from the HTTP connection and not just once the segment transmission ends. Emulations that do not have this feature cannot be used for detecting video stalls accurately as they may find stalls where there are not.

Although in the literature we find many ABR algorithms (a complete survey can be found in [38]), due to the differences on the underlying platform, we have not been able to use nor adapt some popular ABR algorithms like the solutions implemented on "dash.js" web player, such as BOLA [39]. Since it is a web player, its code is based on HTML/JavaScript, whereas ExoPlayer is based on the native Android API. Also, the "dash.js" player has some features that are not present in Exoplayer like segment abandonment. Therefore, all the solutions based on "dash.js" are not directly portable to ExoPlayer since "dash.js" is only browser-oriented and not compatible with the ExoPlayer library.

The adaptation algorithms considered in this work have been tested on different scenarios: 4 channels with constant bandwidth (1, 2, 5 and 10 Mbps) and 4 channels with variable bandwidth. In particular, the first variable channel (staircase) switches between 2, 4, 8 and 4 Mbps in loop every 100 s, the second (stepped) switches between 2 and 8

Mbps every 100 s, whereas the other two are 4G scenarios obtained from traces of field measurements carried out by the Ghent University, specifically a bus and a car in motion, publicly available in [40].

5.1 Objective evaluation

In the objective evaluation, we calculate the main parameters that affect the Quality of Experience (number and duration of stalls, stalling ratio, average representation, and number of representation switches) to calculate the QoE models proposed.

Regarding the bitrate-based QoE models, we have fixed λ=1 and μ=6000, that is, 1 second of rebuffering has the same penalty as the bitrate reduction of a chunk by 6000 kbps. These values are suggested in [42] ([5] uses μ=3000, a value less restrictive regarding rebufferings). As no details about the $q(\cdot)$ function are shown in the Yin et al. QoE model, we have assumed, for simplicity, that $q(x)=x$, which accomplishes the only requirement of being an increasing function. Regarding the PNSR-based QoE model, we have initially fixed ζ=1 and η=3, and with regards to VMAF, we have considered λ=1 and γ=900. We have used these values of η and γ since, according to [43], a stalling ratio of 1% is considered to be noticeable for users, while values higher than 10% are considered to be not acceptable. Analyzing Fig. 1, when the stalling ratio is 10%, for the case of η=3, the QoE_{PSNR} is lower than 10 dB, which is considered unacceptable for users. Following the same criteria, in Fig. 3 we see that the γ that best accomplishes the previous condition is γ=900, since it provides a value of QoE_{VMAF}=0 when the stalling ratio is 10%. Nevertheless, one of the studies presented in this paper evaluates the performance of QoE_{PSNR} and QoE_{VMAF} for different values of η and γ, respectively.

Also, in order to make an accurate comparison regarding the Yin et al. QoE model, in the evaluation, we have not considered the start-up delay since the formula of Yin et al. does not take it into account, so δ=0.

ITU-T P.1203 [8] has also been used to estimate the QoE of the evaluations. This recommendation describes a set of objective parametric quality assessment modules that help to predict the impact of media encoding and observed IP network impairments on quality experienced by the end-user in multimedia streaming applications. This standard has been developed especially for TCP-type streaming like HAS. It takes into account the initial delay and the playback stalls while computing the video quality with NR VQA algorithms. The recommendation includes four different modes of operation, with different complexity both of the input information and the model algorithms. We have used the implementation available in [44] with the extension for the VP9 codec available in [45]. Because of the limitations of the codec extension, the evaluation has been carried out using the ITU-T P.1203 mode 0 since there is not any available ITU-T P.1203 implementation valid for modes 1, 2, or 3 for the VP9 codec. In the evaluation, the videos have been encoded with VP9, the latest open and royalty-free video coding format developed by Google and one of the most used video codecs nowadays.

Three videos have been chosen to perform the evaluation, both created by the Blender Foundation [41]: "Elephants Dream," "Tears of Steel," and a longer video (whose duration is about 46 minutes) composed by 4 open-source videos (the aforementioned videos “Elephants Dream” and “Tears of Steel” as well as the videos “Sintel” and “Big Buck Bunny”). It is worth noting that the chosen videos offer two different kinds of contents: on the one hand, three cartoon

videos, and on the other hand, a sci-fi movie with human actors. In this way, these videos cover a vast spectrum of types of multimedia content available nowadays.

All representations have a Full HD resolution, a frame rate of 24 fps, and a segment size of 10 seconds. Also, different values of CRF (Constant Rate Factor) have been used: from 5 (better quality) to 60 (lower quality) in intervals of 5, that is, a total of 12 video qualities. We have encoded videos with CRFs from 5 to 60 in steps of 5 to be systematic and to better evaluate the performance of the algorithms in terms of representation switches even though some representations are hardly ever selected by the player. This high number of representations can lead to a considerable amount of representation switches when playing the video, but most of them are unnoticeable for users. Table 1 summarizes the main characteristics of the videos used for the evaluation.

Table 1. Characteristics of the videos used in the objective evaluation.

Video	Duration (s)	Number of segments	Resolution	Frame rate (fps)	Codec
Elephants Dream	654	66	1920x1080	24	VP9
Tears of Steel	734	74	1920x1080	24	VP9
Mix (Sintel - Big Buck Bunny - Elephants Dream - Tears of Steel)	2757	276	1920x1080	24	VP9

All video playbacks have been carried out using an instance of the official Android 8 emulator running on HP Pavilion dv6 (i7/6GB) with the Ubuntu 18.04 Linux distribution. Also, on the server side, we have used a local instance of Apache 2.4 to avoid undesired bandwidth limitations.

Finally, to obtain the data shown in the evaluation section, 5 iterations have been carried out for each algorithm, channel, and video under review, providing narrow confidence intervals. Specifically, a total of almost 98 hours (that is, about 4 days) of video have been displayed for the objective evaluation.

5.2 Subjective evaluation

The subjective evaluation has been carried out taking into account the recommendations of the standard ITU-T P.913 [10]. In this case, the only scenario considered is the 4G-car channel since, as we will see, it is the most demanding channel.

The subjective evaluation was carried out in a laboratory of the Universitat Politècnica de València for 3 weeks, in sessions of 60 minutes in length. A maximum of 4 people participated in every session to optimize the evaluation time. The users did not interact among them since, during the evaluation, there were two coordinators in charge of ensuring the correct performance of the test. A total of 24 people (15 men and 9 women) participated in the study, which an age between 20 and 42 years.

The evaluation was carried out using an Apple 24” iMac (model number A1225), with a resolution of 1920x1200, and an aspect ratio of 16:10. The color and luminosity of the screen were configured according to the computer recommendation. The room luminosity was 25 lux, and the distance among the screen and the test subjects was 3 times the height of the screen (3H).

Again, in the subjective evaluation, we have used the videos "Elephants Dream" and "Tears of Steel." It is important to highlight that the duration of the two videos (between 10 and 12 minutes) allows evaluating the QoE of the users

appropriately. In fact, one of the strengths of this paper is that, unlike most works that evaluate the QoE, we use videos with a larger duration, which makes the evaluation process much more time-consuming.

As in the objective evaluation, we have used a segment size of 10 seconds, VP9, a resolution of 1920x1080, 24 fps, and CRF values between 5 and 60. As we detail in the evaluation section, in this case, we have only considered two ABR algorithms (Look Ahead and Müller), thus leading to four cases: case A ("Elephants Dream" video and Look Ahead algorithm), B ("Elephants Dream" and Müller algorithm), C ("Tears of Steel" and Look Ahead) and D ("Tears of Steel" and Müller), as detailed in Table 2.

Table 2. Characteristics of the videos used in the subjective evaluation.

Case	Video	Algorithm	Duration (s)	Number of segments	Resolution	Frame rate (fps)	Codec
A	Elephants Dream	Look Ahead	654	66	1920x1080	24	VP9
B	Elephants Dream	Müller	654	66	1920x1080	24	VP9
C	Tears of Steel	Look Ahead	734	74	1920x1080	24	VP9
D	Tears of Steel	Müller	734	74	1920x1080	24	VP9

As mentioned, there are several rating scales for measuring the MOS. The most commonly used is the 5-point ACR scale: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), and 1 (Bad). This metric is used in the first part of each session. In the second part, when comparing two algorithms, we use the CCR scale, in which a content is compared to a previous content according to the following scale: much better (3), better (2), slightly better (1), about the same (0), slightly worse (-1), worse (-2) and much worse (-3).

Each session was divided into two parts of 30 minutes approximately, with an intermission of 5 minutes. In each part, subjects evaluated one of the two videos (“Elephants Dream” or “Tears of Steel”). Tests subjects carried out two different evaluations in each part: in the first they evaluated each algorithm independently (ACR test); whereas in the second, subjects evaluated the visual quality of an algorithm regarding the other algorithm (CCR test). It is worth noting that the order of the videos and algorithms shown to the users was not the same in all the studies. For example, sometimes it was shown case A first, then case B, and then the other video (case C and then case D, for instance); and others it was shown first case D, then C, and then the other video (case A and then case B, for instance). But always the two cases belonging to the same video are displayed consecutively, for example, the sequence B-C-D-A is not valid, since cases B and A (which are referred to the same video) are not displayed one after the other.

Fig. 4 shows an example of the sequence followed to carry out the subjective evaluation. At the beginning of the evaluation, a visual acuity test was performed to check the validity of the evaluation carried out by each user. Finally, at the end of the test, each user was asked about ordering from the most to the least representative metric related to their user experience among a list of four relevant QoE parameters: stalls, start-up delay, quality of the video displayed and number of representation switches.

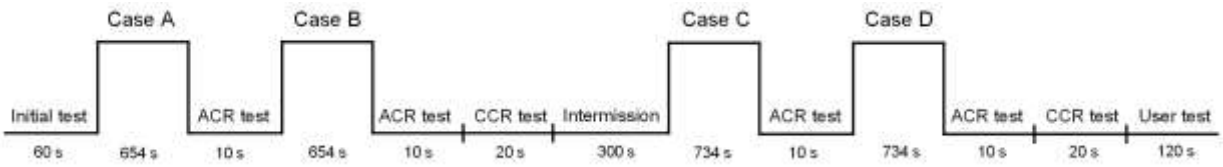
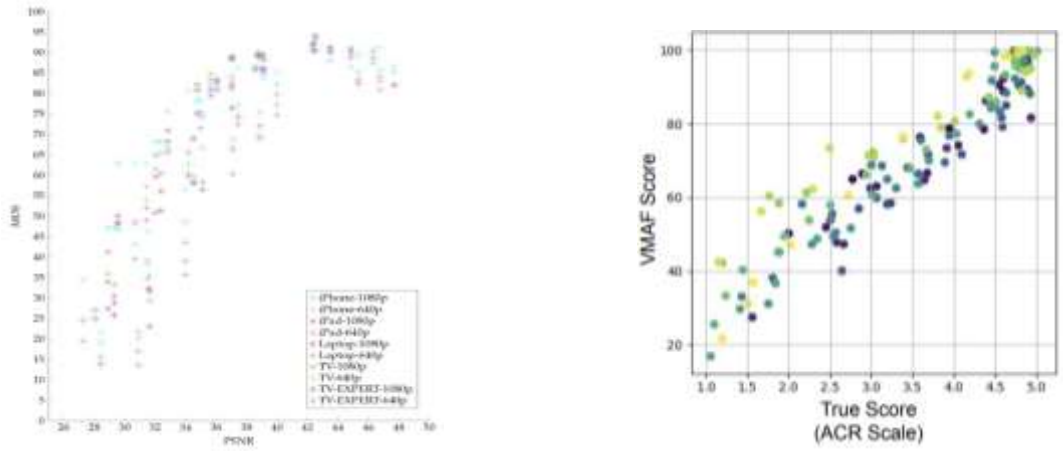


Fig. 4 Sequence of the subjective evaluation.

The subjective study helps to evaluate the performance of the proposed QoE objective models. To that extent, it is important to try to establish a relationship between the MOS and the QoE models proposed. In this sense, in the literature, we can find few works which study the relationship of the MOS an objective metrics such as PSNR, SSIM, VQM, or VMAF. Fig. 5-a shows the relationship between MOS and PSNR extracted from a study carried out by the University of Waterloo [46] using different videos and display devices. Note that the y-axis of Fig. 5-a can be translated to ACR scale considering that the highest value (100) corresponds to ACR=5, and the lowest value (0) corresponds to ACR=1. In the figure we can see a point cloud as a consequence of the different display screen sizes, but, in general, the MOS increases as PSNR does. On the other hand, among the collection of studies carried out by Netflix about VMAF, [47] presents a mapping between ACR scale to VMAF scale, shown in Fig. 5-b. We can see that a fair QoE is obtained, in general, when VMAF is higher than 60; a good QoE is obtained when VMAF is higher than 80; and when VMAF is close 100 we get an excellent QoE.



(a) MOS vs. PSNR [46]. **(b)** MOS (ACR) vs. VMAF [47].

Fig. 5 Relationship between MOS and objective metrics.

6. Evaluation

6.1 Objective evaluation

In this section, we evaluate the main parameters that affect the QoE when watching a video: number and duration of stalls, average representation, and number of representation switches. This information is shown in Table 3 and Table 5 for the scenarios and algorithms under study and for the videos "Elephants Dream" and "Tears of Steel," respectively. These parameters are used to calculate the different QoE models analyzed in this work, which are shown in Table 4 and Table 6 for the videos "Elephants Dream" and "Tears of Steel," respectively. All the tables present average results obtained from the 5 iterations carried out in each scenario. The tables also show the difference (in % or absolute value) regarding the algorithm that provides the best value in each scenario.

Considering the information of the first video shown in Table 3, we see that Look Ahead outperforms SARA and Müller algorithms in terms of stalls since, according to the table, Look Ahead does not have stalls whereas both SARA and Müller suffer from stalls in the most demanding channels (1 Mbps, 2 Mbps, and the two 4G channels). The table

reflects that the lack of stalls does not imply a meaningful decrease in the average representation regarding the Look Ahead algorithm. In fact, the average representation obtained by Look Ahead for all the scenarios under consideration is slightly lower than the average representation of the best case (SARA or Müller, depending on the scenario). As regards the number of representation switches, as shown in Table 3, Look Ahead is the algorithm that offers, in general, the best values.

Table 3. Evaluation of the video "Elephants Dream" for different algorithms and channels.

Channel	Adaptation algorithm	Number of stalls	Stalls duration (s)	Stalling ratio	Average repres. [0-11]		Representation switches	
1 Mbps	Müller	1.00	5.81	0.89%	3.64	-0.42%	49.00	+9.38%
	SARA	1.00	5.98	0.91%	3.66	max	49.20	+9.82%
	Look Ahead	0.00	0.00	0.00%	3.27	-10.49%	44.80	min
2 Mbps	Müller	1.00	5.48	0.84%	3.64	-0.42%	49.20	+8.85%
	SARA	1.00	5.95	0.91%	3.66	max	49.20	+8.85%
	Look Ahead	0.00	0.00	0.00%	3.30	-9.62%	45.20	min
5 Mbps	Müller	0.00	0.00	0.00%	7.25	-2.27%	45.20	min
	SARA	0.00	0.00	0.00%	7.41	max	48.20	+6.64%
	Look Ahead	0.00	0.00	0.00%	6.56	-11.59%	47.40	+4.87%
10 Mbps	Müller	0.00	0.00	0.00%	8.66	-1.57%	42.40	+13.98%
	SARA	0.00	0.00	0.00%	8.79	max	40.40	+8.60%
	Look Ahead	0.00	0.00	0.00%	8.17	-7.06%	37.20	min
2-4-8-4 Mbps	Müller	0.00	0.00	0.00%	6.48	max	44.60	min
	SARA	0.00	0.00	0.00%	6.17	-4.77%	51.60	+15.70%
	Look Ahead	0.00	0.00	0.00%	6.07	-6.37%	46.40	+4.04%
4G-bus	Müller	0.60	9.49	1.45%	7.93	-9.24%	43.80	+25.14%
	SARA	0.40	2.58	0.39%	8.06	-7.81%	45.60	+30.29%
	Look Ahead	0.00	0.00	0.00%	8.74	max	35.00	min
4G-car	Müller	0.80	13.00	1.99%	8.74	-3.90%	37.80	min
	SARA	2.00	21.46	3.28%	9.10	max	38.60	+2.12%
	Look Ahead	0.00	0.00	0.00%	8.53	-6.19%	38.00	+0.53%

Table 4. Evaluation of the QoE models (λ=1, μ=6000, ζ=1, η=3, γ=900) for the video "Elephants Dream."

Channel	Adaptation algorithm	QoE by Yin et al. (M)		QoE modified (M)		QoE PSNR (dB)		QoE VMAF		QoE P.1203	
1 Mbps	Müller	40.03	-41.26%	11.01	-73.89%	31.02	-7.83	70.43	-7.65%	4.24	-0.34
	SARA	37.69	-44.69%	9.94	-76.42%	30.81	-8.04	70.95	-6.97%	4.24	-0.34
	Look Ahead	68.15	max	42.17	max	38.84	max	76.26	max	4.58	max
2 Mbps	Müller	40.10	-42.50%	12.62	-70.44%	31.36	-7.52	70.88	-7.41%	4.24	-0.34
	SARA	36.27	-47.99%	9.80	-77.06%	30.83	-8.06	70.32	-8.14%	4.24	-0.34
	Look Ahead	69.74	max	42.71	max	38.88	max	76.55	max	4.58	max
5 Mbps	Müller	340.19	-3.46%	186.74	-6.51%	44.57	-0.42	91.99	-1.06%	4.61	-0.01
	SARA	352.39	max	199.75	max	44.99	max	92.98	max	4.61	max
	Look Ahead	241.16	-31.56%	162.47	-18.66%	43.67	-1.32	90.68	-2.48%	4.60	-0.01
10 Mbps	Müller	560.55	-3.42%	337.25	-1.22%	47.18	-0.09	94.27	-0.89%	4.63	-0.01
	SARA	580.38	max	341.43	max	47.27	max	95.11	max	4.64	max
	Look Ahead	464.00	-20.05%	289.87	-15.10%	46.46	-0.80	93.92	-1.26%	4.63	-0.01
2-4-8-4 Mbps	Müller	288.99	max	122.93	-6.48%	43.56	max	86.99	-0.07%	4.60	max
	SARA	217.97	-24.58%	117.41	-10.69%	42.07	-1.49	85.71	-1.53%	4.60	max
	Look Ahead	249.27	-13.74%	131.46	max	42.86	-0.70	87.04	max	4.60	max
4G-bus	Müller	377.47	-38.83%	174.44	-52.02%	33.98	-13.67	78.04	-16.30%	4.40	-0.22
	SARA	402.49	-34.77%	225.37	-38.02%	40.92	-6.73	89.13	-4.40%	4.49	-0.14
	Look Ahead	617.04	max	363.60	max	47.65	max	93.24	max	4.62	max
4G-car	Müller	548.47	-0.02%	311.10	-14.31%	32.92	-14.36	72.95	-22.37%	4.23	-0.40
	SARA	499.57	-8.94%	274.16	-24.49%	28.34	-18.93	61.95	-34.07%	3.97	-0.66
	Look Ahead	548.60	max	363.07	max	47.27	max	93.97	max	4.63	max
AVG.	Müller	313.69	-2.75%	165.16	-17.15%	37.80	-5.87	80.79	-7.54%	4.42	-0.18
	SARA	303.82	-5.81%	168.27	-15.59%	37.89	-5.77	80.88	-7.44%	4.40	-0.21
	Look Ahead	322.57	max	199.33	max	43.66	max	87.38	max	4.61	max

Table 5. Evaluation of video "Tears of Steel" for different algorithms and channels.

Channel	Adaptation algorithm	Number of stalls	Stalls duration (s)	Stalling ratio	Average repres. [0-11]		Representation switches	
1 Mbps	Müller	0.00	0.00	0.00%	3.28	-0.08%	54.00	+3.85%
	SARA	0.00	0.00	0.00%	3.28	max	52.40	+0.77%
	Look Ahead	0.00	0.00	0.00%	2.78	-15.36%	52.00	min
2 Mbps	Müller	0.00	0.00	0.00%	3.28	-0.17%	54.60	+3.80%
	SARA	0.00	0.00	0.00%	3.29	max	53.80	+2.28%
	Look Ahead	0.00	0.00	0.00%	2.77	-15.60%	52.60	min
5 Mbps	Müller	0.00	0.00	0.00%	6.62	max	50.80	+16.51%
	SARA	0.00	0.00	0.00%	6.61	-0.08%	54.40	+24.77%
	Look Ahead	0.00	0.00	0.00%	6.07	-8.32%	43.60	min
10 Mbps	Müller	0.00	0.00	0.00%	7.94	-0.07%	46.20	min
	SARA	0.00	0.00	0.00%	7.95	max	48.20	+4.33%
	Look Ahead	0.00	0.00	0.00%	7.09	-10.80%	48.60	+5.19%
2-4-8-4 Mbps	Müller	0.00	0.00	0.00%	6.06	max	54.20	+18.86%
	SARA	0.00	0.00	0.00%	5.78	-4.51%	53.40	+17.11%
	Look Ahead	0.00	0.00	0.00%	5.52	-8.88%	45.60	min
4G-bus	Müller	0.00	0.00	0.00%	8.20	-2.93%	51.80	+17.19%
	SARA	0.00	0.00	0.00%	8.45	max	44.20	min
	Look Ahead	0.00	0.00	0.00%	8.02	-5.06%	53.20	+20.36%
4G-car	Müller	0.80	7.43	1.01%	8.17	-0.79%	43.20	min
	SARA	2.60	30.04	4.09%	8.23	max	43.80	+1.39%
	Look Ahead	0.00	0.00	0.00%	7.99	-2.92%	50.20	+16.20%

Table 6. Evaluation of the QoE models (λ=1, μ=6000, ζ=1, η=3, γ=900) for the video "Tears of Steel."

Channel	Adaptation algorithm	QoE by Yin et al. (M)		QoE modified (M)		QoE PSNR (dB)		QoE VMAF		QoE P.1203	
1 Mbps	Müller	52.15	-3.32%	53.97	-1.75%	36.79	max	82.23	-1.15%	4.58	-0.01
	SARA	53.94	max	54.93	max	36.78	-0.01	83.19	max	4.58	-0.01
	Look Ahead	42.88	-20.52%	47.84	-12.92%	36.13	-0.66	79.47	-4.47%	4.59	max
2 Mbps	Müller	51.19	-1.47%	54.01	-0.05%	36.68	-0.07	82.63	max	4.58	-0.01
	SARA	51.96	max	54.03	max	36.75	max	82.23	-0.49%	4.58	-0.01
	Look Ahead	42.37	-18.46%	47.76	-11.61%	36.10	-0.64	79.47	-3.82%	4.59	max
5 Mbps	Müller	220.29	max	226.07	-0.33%	41.16	max	94.68	-0.09%	4.64	max
	SARA	216.40	-1.77%	226.83	max	41.11	-0.04	94.76	max	4.64	max
	Look Ahead	197.01	-10.57%	203.41	-10.32%	40.74	-0.42	94.06	-0.73%	4.62	-0.02
10 Mbps	Müller	442.57	max	398.68	-1.45%	42.85	-0.06	95.97	-0.53%	4.64	-0.06
	SARA	427.21	-3.47%	404.53	max	42.91	max	96.48	max	4.70	max
	Look Ahead	298.48	-32.56%	320.08	-20.88%	41.82	-1.09	95.33	-1.19%	4.64	-0.06
2-4-8-4 Mbps	Müller	171.80	max	199.42	max	40.33	max	92.69	max	4.61	max
	SARA	153.78	-10.49%	183.75	-7.86%	39.80	-0.53	90.59	-2.26%	4.60	-0.01
	Look Ahead	154.88	-9.85%	172.30	-13.60%	39.91	-0.42	91.57	-1.21%	4.60	-0.01
4G-bus	Müller	441.90	-18.96%	410.85	-19.04%	42.77	-0.80	95.96	-0.94%	4.64	-0.01
	SARA	545.28	max	507.49	max	43.57	max	96.87	max	4.65	max
	Look Ahead	375.31	-31.17%	425.60	-16.14%	42.76	-0.81	95.94	-0.96%	4.65	max
4G-car	Müller	550.97	max	508.80	-0.82%	33.89	-8.99	83.03	-13.21%	4.26	max
	SARA	399.78	-27.44%	384.93	-24.97%	21.52	-21.36	55.10	-42.41%	3.77	-0.50
	Look Ahead	453.31	-17.72%	513.02	max	42.88	max	95.67	max	4.11	-0.15
AVG.	Müller	275.84	max	264.54	max	39.21	-0.84	89.60	-0.69%	4.56	max
	SARA	264.05	-4.27%	259.50	-1.91%	37.49	-2.56	85.60	-5.11%	4.50	-0.06
	Look Ahead	223.46	-18.99%	247.14	-6.58%	40.05	max	90.22	max	4.54	-0.02

Analyzing now the values of QoE presented in Table 4, we see that the five QoE models offer coherent results regarding the 1 and 2 Mbps channel: the algorithm that provides the highest value of QoE is Look Ahead, that is, the only algorithm that does not have stalls. Also, it is the algorithm with the lowest number of representation switches and, although it has the lowest average representation, the difference regarding the best case is about only 10%. Similar conclusions arise in the two 4G channels. However, in the particular case of the 4G-car, regarding the QoE model proposed by Yin et al., we find an apparently incoherent result, since the QoE of Look Ahead and Müller is almost

the same even though: the latter algorithm has an average stalling ratio of 2% with an average duration of 13 seconds, the average representation is hardly better in Müller than Look Ahead (less than a 5%) and both algorithms have the same average number of representation switches. Finally, in the least demanding channels (5, 10, and 2-4-8-4 Mbps), as no stalls occur, the best QoE is obtained by the algorithm that provides the best average representation. The table also shows the average values of the 7 scenarios under consideration for each adaptation algorithm, and we can see that Look Ahead is the algorithm that provides the best values in all QoE models.

Regarding the video "Tears of Steel," according to Table 5, the only scenario where there are stalls is the 4G-car. In that channel, SARA has a stalling ratio of 4% with an average stall duration of 30 seconds, Müller has a stalling ratio of 1% and 7.4 seconds of average stall duration whereas Look Ahead does not have stalls. Analyzing the QoE models shown in Table 6 referred to the 4G-car channel, we see that both the QoE by Yin et al. and the ITU-T P.1203 offer apparently incoherent values since they provide the best QoE for the case of the Müller algorithm instead of Look Ahead. This seems contradictory taking into account that Look Ahead does not suffer from stalls, and the average representation of Müller is just 2.25% higher than the average representation of Look Ahead. In the rest of the scenarios, all QoE models offer reasonable values, since usually the algorithm with the best values of QoE is the algorithm with the highest average representation, as long as there is not a meaningful difference regarding representation switches. Checking the average values, we see that Look Ahead provides the best value for the PSNR and VMAF-based QoE models, whereas Müller has the highest QoE in the Yin et al.-based model and ITU-T P.1203.

As a summary, we can say that, according to the results, the proposed QoE model modified outperforms the original QoE model by Yin et al., which offers incoherent results in some scenarios. Although it is true that the QoE model by Yin et al. allows assigning more weight to the stalls (with the parameter μ) and thus the results obtained could change, in this work we have used the most restrictive value among the μ suggested by [5] and [42]. Moreover, the two proposed PSNR and VMAF-based QoE models offer consistent results in all scenarios. Finally, regarding the ITU-T P.1203, we see that it offers rather optimistic results in many cases (the lowest value of the QoE is 3.97 in "Elephants Dream" and 3.77 in "Tears of Steel"). This can be because we have considered the simplest mode of operation of the standard due to the limitations of the VP9 codec for the standard. Also, we see that the proposed algorithm Look Ahead outperforms Müller and SARA ABR algorithms in terms of the number and duration of video playback stalls, without hardly decreasing the average video quality, therefore Look Ahead provides a better Quality of Experience.

The results for the mixed video are shown in Table 7 and Table 8. In this video, all the algorithms, even Look Ahead, cause stalls during the display in the most exigent channels (especially the 4G-car and the stepped channel –8-2 Mbps–). In any case, again, it is Look Ahead the algorithm that offers the best results in terms of number and duration of stalls without hardly getting worse the average representation and with a similar number of representation switches, as the average values of the five scenarios show.

For the analysis of the QoE models, we examine the particular case of the 4G-car. In that channel, on average, Müller has 5.60 stalls whose duration is 63.19 seconds, SARA has 9.2 stalls of average duration 100.44 seconds, whereas Look Ahead has 4 stalls with a total duration of 24.37 seconds. The average representation is slightly higher in SARA (8.87) than in Look Ahead (8.81) and Müller (8.66), and the number of representation switches is quite similar in both algorithms (between 161.20 and 168.20). Considering this, it seems that Look Ahead behaves better

than SARA and Müller because of the difference in terms of stalls and stalls duration. However, the QoE model proposed by Yin et al. is better in Müller (2227 M) and SARA (2120 M) than in Look Ahead (2110 M). In a real scenario, it is difficult to believe that users perceive a better experience watching a video playback that uses an algorithm that causes more stalls than another that has almost the same average quality and much fewer stalls. In contrast, the Yin et al. QoE modified model shows a completely different result since, in this case, the QoE of Look Ahead (1610 M) is better than the QoE of the SARA algorithm (1113 M) and the Müller algorithm (1215 M). Likewise, this result is coherent both with the PSNR-based and the VMAF-based QoE models. Thus, QoE_{PSNR} for Look Ahead (35.74 dB) is much better than QoE_{PSNR} for Müller (25.50 dB) and SARA (19.72 dB), whereas QoE_{VMAF} for Look Ahead (88.37) is also better than QoE_{VMAF} for Müller (72.72) and SARA (60.74). Finally, the ITU-T P.1203 provides coherent results but not so noticeable as those provided by the models proposed in this paper.

Table 7. Evaluation of the mixed video for different algorithms and channels.

Channel	Adaptation algorithm	Number of stalls	Stalls duration (s)	Stalling ratio	Average repres. [0-11]		Representation switches	
1 Mbps	Müller	1.00	6.86	0.25%	3.62	max	194.80	+1.4%
	SARA	1.00	9.05	0.33%	3.55	-1.9%	208.40	+8.4%
	Look Ahead	0.00	0.00	0.00%	3.17	-12.4%	192.20	min
10 Mbps	Müller	0.00	0.00	0.00%	8.74	max	166.20	min
	SARA	0.00	0.00	0.00%	8.13	-7.0%	192.00	+15.5%
	Look Ahead	0.00	0.00	0.00%	7.87	-10.0%	174.80	+5.2%
8-2 Mbps	Müller	1.80	22.48	0.82%	9.16	max	173.40	min
	SARA	4.60	46.91	1.70%	8.23	-10.2%	185.60	+7.0%
	Look Ahead	1.00	3.71	0.13%	7.87	-14.1%	174.60	+0.7%
4G-bus	Müller	0.00	0.00	0.00%	8.74	-3.2%	197.80	+14.2%
	SARA	2.40	42.44	1.54%	9.03	max	173.20	min
	Look Ahead	0.00	0.00	0.00%	8.69	-3.8%	174.20	+0.6%
4G-car	Müller	5.60	63.19	2.29%	8.66	-2.4%	168.20	+4.3%
	SARA	9.20	100.44	3.64%	8.87	max	161.20	min
	Look Ahead	4.00	24.37	0.88%	8.81	-0.7%	163.40	+1.4%

Table 8. Evaluation of the QoE models (λ=1, μ=6000, ζ=1, η=3, γ=900) for the mixed video.

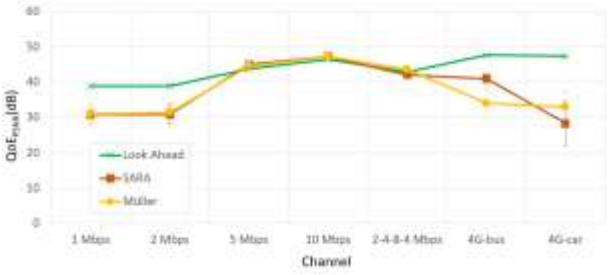
Channel	Adaptation algorithm	QoE by Yin et al. (M)		QoE modified (M)		QoE PSNR (dB)		QoE VMAF		QoE P.1203	
1 Mbps	Müller	192.07	max	166.72	-13.8%	35.21	-3.20	80.67	max	4.13	-0.46
	SARA	152.46	-20.6%	144.06	-25.5%	34.01	-4.40	79.74	-1.2%	4.13	-0.46
	Look Ahead	176.27	-8.2%	193.43	max	38.41	max	80.58	-0.1%	4.59	max
10 Mbps	Müller	2133.48	max	1473.66	max	46.70	max	96.81	max	4.65	-0.01
	SARA	1559.89	-26.9%	1152.67	-21.8%	45.74	-0.96	96.11	-0.7%	4.66	max
	Look Ahead	1443.09	-32.4%	1212.40	-17.7%	45.35	-1.35	95.91	-0.9%	4.65	-0.01
8-2 Mbps	Müller	2481.90	max	1650.38	max	36.88	-6.28	88.65	-5.9%	4.07	-0.09
	SARA	1403.60	-43.4%	916.80	-44.4%	28.27	-14.89	79.17	-16.0%	3.52	-0.64
	Look Ahead	1432.60	-42.3%	1201.11	-27.2%	43.16	max	94.20	max	4.16	max
4G-bus	Müller	1956.69	-9.2%	1229.98	-21.1%	46.26	-0.34	95.67	-1.1%	4.65	max
	SARA	2154.50	max	1354.25	-13.1%	30.69	-15.91	81.74	-15.5%	3.93	-0.72
	Look Ahead	2010.71	-6.7%	1558.94	max	46.60	max	96.74	max	4.65	max
4G-car	Müller	2227.83	max	1214.99	-24.6%	25.50	-10.24	72.72	-17.7%	3.37	-0.08
	SARA	2119.87	-4.8%	1113.45	-30.9%	19.72	-16.02	60.74	-31.3%	3.03	-0.42
	Look Ahead	2100.26	-5.7%	1610.78	max	35.74	max	88.37	max	3.45	max
AVG.	Müller	1737.43	max	1156.14	-6.10%	38.95	-1.88	88.12	-2.8%	4.17	max
	SARA	1566.84	-9.8%	1040.03	-15.5%	34.17	-6.66	82.37	-9.2%	3.99	-0.19
	Look Ahead	1543.87	-11.1%	1231.24	max	40.83	max	90.70	max	4.16	-0.01

Similar conclusions arise when analyzing other scenarios, for example, the stepped 8-2 Mbps channel. In that channel, despite the duration of the stalls when video is displayed using the SARA and Müller algorithms, it is the

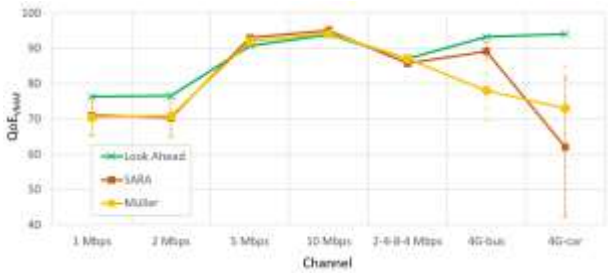
later algorithm which provides, by far, the best values of QoE according to the model defined by Yin et al. In contrast, both PSNR and VMAF-based QoE model and ITU-T P.1203 indicate that Look Ahead provides the best QoE.

Henceforth, to simplify the analysis, we only consider the proposed PSNR and VMAF-based QoE models, apart from the standard ITU-T P.1203.

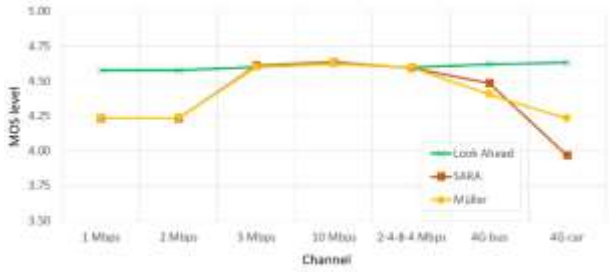
In this sense, Fig. 6 show the evaluation carried out for the video “Elephants Dream” and the algorithms under study for the PSNR-based QoE model, VMAF-based QoE model, and ITU-T P.1203 QoE model. We can see, now visually, how the three models offer very similar results in all scenarios, providing the lowest QoE values in the most demanding channels. Regarding Fig. 6-a (QoE_{PSNR} model), the values shown in the figure have been obtained by fixing ζ=1, η=3, δ=0. Also, each algorithm, for each channel, contains a lower error bar that represents the value obtained when η=4 (a higher penalty for stalls) and an upper error bar that represents the value obtained when η=2. Similar occurs in Fig. 6-b (QoE_{VMAF} model), where the lower error bar represents the value obtained when γ=1500 and the upper error bar is the case when γ=300. Analyzing the three figures, we can see how the three models behave rather similar, although the ITU-T P.1203 offers quite optimistic results.



(a) QoE_{PSNR} (ζ=1, η=3, δ=0).



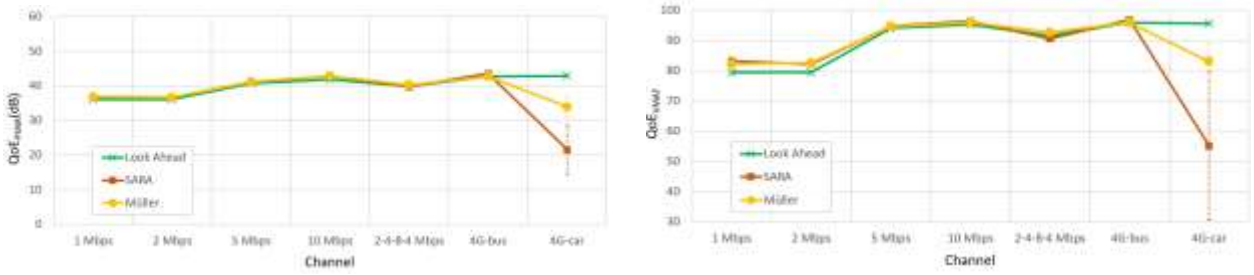
(b) QoE_{VMAF} (λ=1, γ=900, δ=0).



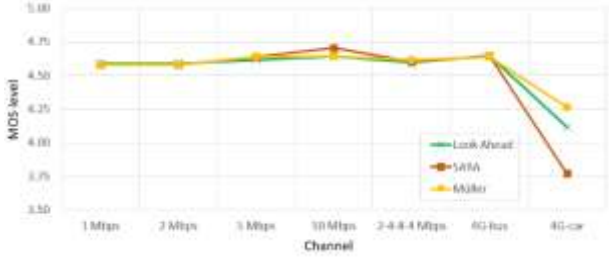
(c) $QoE_{ITU\text{-}T\ P.1023}$.

Fig. 6 QoE of the video “Elephants Dream.”

Similar results appear when analyzing the video "Tears of Steel," which evaluation is shown in Fig. 7. The only remarkable difference is the value of the Look Ahead algorithm in the 4G-car channel in the ITU-T P.1203 QoE model. As mentioned when analyzing previous tables, this value seems apparently incoherent taking into account the absence of stalls of Look Ahead and that the average representation obtained using this algorithm is only 2.92% lower than the best case.



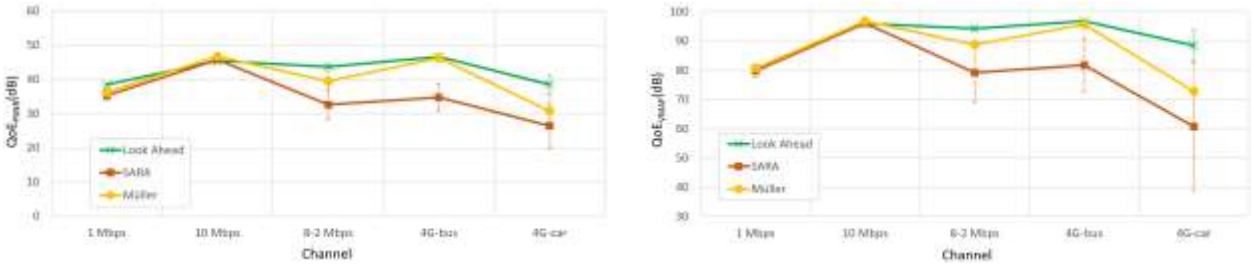
(a) QoE$_{PSNR}$ (ζ=1, η=3, δ=0). **(b)** QoE$_{VMAF}$ (λ=1, γ=900, δ=0).



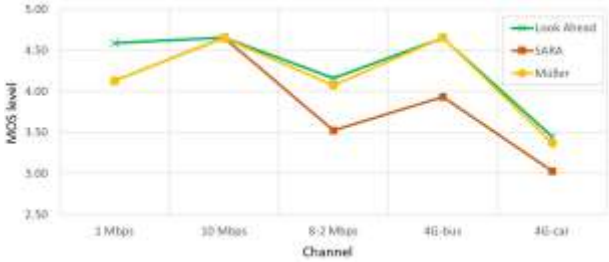
(c) QoE$_{ITU\text{-}T\ P.1023}$.

Fig. 7 QoE of the video "Tears of Steel."

Finally, Fig. 8 depicts that the three QoE model behave rather similar for the mixed video, although the ITU-T P.1203 model penalizes considerably the number of stalls for the 4G-car scenario.



(a) QoE$_{PSNR}$ (ζ=1, η=3, δ=0). **(b)** QoE$_{VMAF}$ (λ=1, γ=900, δ=0).



(c) QoE$_{ITU\text{-}T\ P.1023}$.

Fig. 8 QoE of the mixed video.

6.2 Subjective evaluation

Once analyzed the QoE models in the different scenarios, we carry out a subjective evaluation in a particular channel. Specifically, we consider the most demanding scenario among the analyzed channels, the 4G-car channel, and the two DASH adaptation algorithms that behave better according to the previous results: Look Ahead and Müller.

Thus, to simplify, in this section, we consider 4 different cases, as shown in the x-axis of Fig. 9: case A ("Elephants Dream" video and Look Ahead algorithm), B ("Elephants Dream" and Müller), C ("Tears of Steel" and Look Ahead), and D ("Tears of Steel" and Müller).

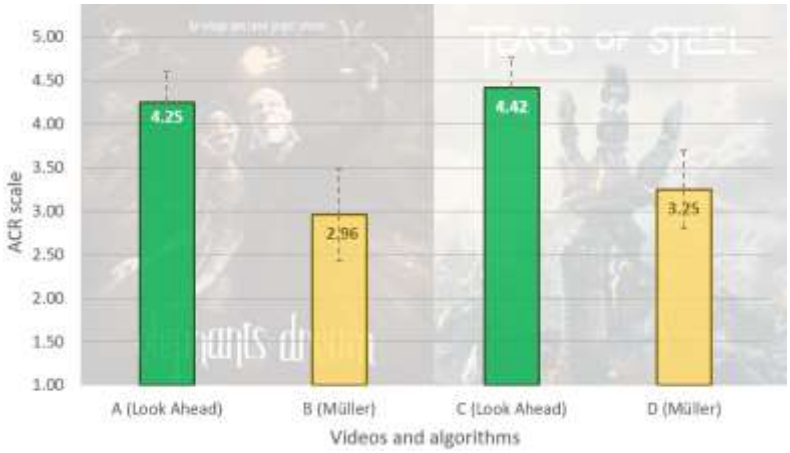


Fig. 9 MOS evaluation for the channel "4G-car".

Fig. 9 shows the MOS evaluation (using the ACR scale) of the videos "Elephants Dream" (left side of the graph) and "Tears of Steel" (right side). It is worth highlighting that the figure includes the 99% confidence intervals for each video and algorithm. As the figure depicts, we can consider these confidence intervals rather narrow, with a maximum interval of ±0.53. Regarding "Elephants Dream," we see that the MOS using the Look Ahead and the Müller algorithms is 4.25 and 2.96, respectively. If comparing these data to the results shown in Table 3, Table 4, and Fig. 6, we see that the proposed PSNR and VMAF-based QoE models offer coherent results considering the mapping between MOS-ACR and PSNR/VMAF shown in Fig. 5: 47.27 dB and 32.92 dB regarding QoE_{PSNR}, and 93.97 and 72.95 regarding QoE_{VMAF} for Look Ahead and Müller algorithms, respectively. On the other hand, although the estimated $QoE_{ITU\text{-}T\ P.1203}$ of Look Ahead (4.63) is rather close to the obtained with the subjective tests (4.25), the value of $QoE_{ITU\text{-}T\_P.1203}$ of Müller (4.23) differs a little bit (a difference of 1.27).

The same conclusions arise when analyzing the video "Tears of Steel." In this case, both the MOS of cases C (4.42) and D (3.25) slightly increases compared to the other video. Regarding QoE_{PSNR}, in this case, $QoE_{PSNR}(C)$ is lower than in the previous video (42.88 dB), and $QoE_{PSNR}(D)$ is slightly higher (33.89 dB). At this point, it is worth mentioning that the PSNR depends on the content of each video, as shown in Fig. 5-a, so this could be the cause of this apparently incoherent value of $QoE_{PSNR}(C)$ when comparing to $QoE_{PSNR}(A)$. Conversely, the VMAF-based QoE model offers the most coherent results (considering Fig. 5-b): $QoE_{VMAF}(C)$ slightly increases (95.67) whereas $QoE_{VMAF}(D)$ increases (83.03), but it is still far from $QoE_{PSNR}(C)$. Finally, the values obtained from the standard ITU-T P.1203 are not in line with the results shown in Fig. 9, since $QoE_{ITU\text{-}T\ P.1203}(C)=4.11$ is worse than $QoE_{ITU\text{-}T\_P.1203}(D)=4.26$.

Therefore, according to the results, although the three analyzed QoE models offer coherent results, the QoE model most similar to human behavior is the VMAF-based QoE model. Nevertheless, it is worth highlighting that the evaluations carried out with the standard ITU-T P.1203 have used the least complex mode of operation, so the results would improve if a more introspective mode were used.

Table 9 summarizes the values obtained in the 4 analyzed cases for the "4G-car" channel. It is important to mention that the proposed QoE_{PSNR} and QoE_{VMAF} models can be refined by adjusting parameters ζ, η, δ, λ, and γ. In this sense, the table shows the QoE under different values of η and γ.

Table 9. Comparison of the subjective evaluation with objective QoE models (λ=1, ζ=1, δ=0) for the "4G-car" channel.

Case	MOS	ITU-T P.1203	QoE_{PSNR} (dB)			QoE_{VMAF} [0-100]		
			η=2	η=3	η=4	γ=300	γ=900	γ=1500
A	4.25	4.63	47.27	47.27	47.27	93.97	93.97	93.97
B	2.96	4.23	37.67	32.92	28.17	84.88	72.95	61.02
C	4.42	4.11	42.88	42.88	42.88	95.67	95.67	95.67
D	3.25	4.26	36.93	33.89	30.86	89.10	83.03	76.96

Moreover, Fig. 10 shows the results of the study carried out for the two videos and algorithms under study. Specifically, the figure shows a subjective comparison among the videos displayed using Müller compared to a reproduction of the same videos but using Look Ahead. In the y-axis, it is shown the CCR scale (among -3 and 3) comparing the reproduction using Müller regarding Look Ahead. Results show that none of the 24 people that made the tests considered that the reproduction using the Müller algorithm was better than that carried out using the Look Ahead algorithm for the two videos.

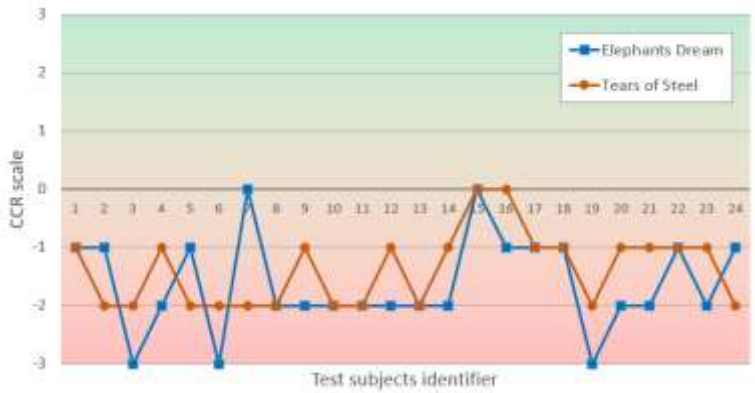


Fig. 10 Comparison of the Müller algorithm regarding Look Ahead for the "4G-car" channel.

Finally, it is presented the results of the study about the question: "As on-demand video user, which of the following metrics is more relevant for you (order from the highest to the lowest priority): Watching the video with a good quality; Not having stalls (or very few) during the playback; Not appreciating meaningful quality switches; Or having a low start-up delay?". The users gave more priority to not having stalls, then to the quality of the video displayed, then to the number of representation switches and finally to the start-up delay. Specifically, 70.83% of test subjects considered that stalls are the most relevant metric when evaluating the quality of video reproduction. This percentage increases to 87.50% for users that believe that stalls are one of the two most relevant metrics. Any user considered the existence of stalls as the least relevant metric. Moreover, 2 out of 3 users considered that the quality of the video displayed is one of the two most relevant metrics (16.67% considered the quality as the most relevant aspect). On the other side, we find the number of representation switches and the initial delay. The first is chosen as the third relevant metric by the 54.17% of users and the least relevant metric by the 20.83%, whereas a 62.50% of users considered that the start-

up delay is the least relevant metric regarding the four metrics under study (although it is the most relevant metric for 1 out of 8 users).

7. Conclusion

This paper has presented three new models for calculating the QoE in an objective way. Both objective and subjective evaluations presented in this work have proved that all three the proposed QoE model by Yin et al. modified, the PSNR-based QoE model and the VMAF-based QoE model offer more realistic results in terms of Quality of Experience than the QoE model proposed by Yin et al. [5] and the recommendation ITU-T P.1203 [8].

The evaluations have been carried out using two well-known ABR algorithms (Müller and SARA) as well as a proposed algorithm called Look Ahead, which takes into account the variability of the bitrate to calculate the best quality level in order to avoid stalls during the visualization of a video. Results have proved that both the number and duration of video playback stalls (rebuffering) are highly reduced using Look Ahead.

As part of the future work, it is worth noting that the proposed PSNR and VMAF-based QoE models can be improved by considering other parameters that affect the user experience, as the number of stalls. In general, it is more annoying for users having many but short stalls than having few although long stalls [48]. For instance, in video playback, users usually prefer having 1 stall of 10 seconds than 10 stalls of 1 second. Moreover, it is worth analyzing how the number of representation switches affects the QoE perceived by the users, a topic deeply analyzed in [49] and [50]. Additionally, it is intended to compare the proposed models with the ATLAS software [19] and with more complex modes of operation of the ITU-T P.1203. In this sense, the ITU is working into Phase II of the ITU-T P.1203 standard, considering a high number of codecs (AVC, HEVC, and VP9) and higher resolution videos (up to UHD), among other features.

It is important to emphasize that it is possible to check the performance of the proposed Look Ahead algorithm by accessing a dedicated server set up by the authors [51], which includes a publicly available App that contains the developed Look Ahead algorithm integrated into ExoPlayer. The App allows to redo the evaluations presented in this paper using the videos and scenarios analyzed in this work. In addition, authors have made available a program in GitHub [6] to encode videos and obtain the bitrate, the PSNR and the VMAF of each segment for each representation, useful for calculating the QoE models proposed in this paper. In this way, these tools guarantee the reproducibility of the ABR algorithm and the three QoE models proposed in this work, as well as all the results presented.

Acknowledgments

This work is supported by the PAID-10-18 Program and by the R&D Line “Technologies for distribution and processing of multimedia information and QoE” from the Universitat Politècnica de València.

References

[1] Cisco webpage, Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper, Available online: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html, accessed: 2019.

[2] Sandvine webpage, The Global Internet Phenomena Report, Available online: https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf, published: 2018, accessed: Sep. 2019.

[3] R. Pantos and W. May, HTTP Live Streaming, ITF RFC vol. 8216, 2017.

[4] ISO/IEC 23009-1, Dynamic adaptive streaming over HTTP (DASH) - Part 1: media presentation description and segment formats, 2012.

[5] X. Yin, V. Sekar, and B. Sinopoli, Toward a principled framework to design dynamic adaptive streaming algorithms over HTTP, in Proc. of the 13th ACM Workshop on Hot Topics in Networks (HotNets), Los Angeles, CA, USA, Oct. 2014, 1-7.

[6] GitHub website, Dashgen. Multimedia Communications Group, available online: https://github.com/comm-iteam/dashgen, accessed: Sep. 2019.

[7] N. Barman and M. G. Martini, QoE modeling for HTTP adaptive video streaming – A survey and open challenges, IEEE Access, 7 (2019) 30831-30859.

[8] International Telecommunication Union (ITU-T), Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport, Recommendation ITU-T P.1203, 2017.

[9] International Telecommunication Union, Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures. BT Series, Broadcasting service, 2012.

[10] International Telecommunication Union, Recommendation ITU-T P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment, 2016.

[11] International Telecommunication Union, Recommendation ITU-T P.910: Subjective Video Quality Assessment methods for multimedia applications, 2008.

[12] International Telecommunication Union, Recommendation ITU-T P.911: Subjective audiovisual quality assessment methods for multimedia applications, 1998.

[13] 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 (TOMM), 14-29 (2018) article no. 29.

[14] 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, 61-4 (2015) 651-665.

[15] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, Objective video quality assessment methods: A classification, review, and performance comparison, IEEE Transactions on Broadcasting, 57-2 (2011) 165-182.

[16] B. Ciubotaru, G. M. Muntean, and G. Ghinea, Objective assessment of region of interest-aware adaptive multimedia streaming quality, IEEE Transactions on Broadcasting, 55-2 (2019) 202-212.

[17] S. Winkler, A. Sharma, and D. Mcnally, Perceptual video quality and blockiness metrics for multimedia streaming applications, in Proc. of the Int. Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark, Sep. 2001, pp. 547-552.

[18] C. G. Bampis and A. C. Bovik, Feature-based prediction of streaming video QoE: Distortions, stalling and memory, Signal Processing: Image Communication, 68 (2018) 218-228.

[19] R. Soundararajan and A. C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Transactions on Circuits and Systems for Video Technology, 23-4 (2013) 684-694.

[20] 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 Proc. of Int. Conf. on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, May 2017.

[21] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software, in Proc. of the 9th ACM Multimedia Systems Conference, Amsterdam, Netherlands, Jun. 2018, pp. 466-471.

[22] X. Deng, L. Chen, F. Wang, Z. Fei, W. Bai, C. Chi. G. Han, and L. Wan, A novel strategy to evaluate QoE for video service delivered over HTTP adaptive streaming, in Proc. of the IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Vancouver, BC, Canada, Sep. 2014.

[23] D. Z. Rodríguez, R. L. Rosa, E. C. Alfaia, J. I. Abrahão, and G. Bressan, Video quality metric for streaming service using DASH standard, IEEE Transactions on Broadcasting, 62-3 (2016) 628-639.

[24] 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, 61-4 (2015) 651-665.

[25] M. N. Garcia, W. Robitza, and A. Raake, On the accuracy of short term quality models for long-term quality prediction, in Proc. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX), Pylos, Greece, May 2015.

[26] 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, 11-1 (2017) 154-166.

[27] C. G. Bampis, Z. Li, and A. C. Bovik, Continuous prediction of streaming video QoE using dynamic networks, IEEE Signal Processing Letters, 24-7 (2017) 1083-1087.

[28] C. G. Bampis and A. C. Bovik, An augmented autoregressive approach to HTTP video stream quality prediction, available online: https://arxiv.org/abs/1707.02709, 2017.

[29] S. Winkler and P. Mohandas, The evolution of Video Quality Measurement: from PSNR to hybrid metrics, IEEE Transactions on Broadcasting, 54-3 (2008) 660-668.

[30] M. P. Sharabayko, O. G. Ponomarev, and R. I. Chernyak, Intra compression efficiency in VP9 and HEVC, Applied Mathematical Sciences, 7-137 (2013) 6803-6824.

[31] Medium webpage, “Toward a practical perceptual video quality metric,” available online: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, published: 2016, accessed: Sep. 2019.

[32] C. G. Bampis, Z. Li, and A. C. Bovik, SpatioTemporal feature integration and model fusion for full reference video quality assessment, IEEE Trans. on Circuits and Systems for Video Technology, 29-8 (2019) 2256-2270.

[33] H. Sheikh and A. Bovik, Image information and visual quality, IEEE Transactions on Image Processing, 15- 2 (2006) 430-444.

[34] S. Li, F. Zhang, L. Ma, and K. Ngan, Image quality assessment by separately evaluating detail losses and additive impairments, IEEE Transactions on Multimedia, 13-5 (2011) 935-949.

[35] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning, 20-3 (1995) 273-297.

[36] C. Müller, S. Lederer, and C. Timmerer, An evaluation of dynamic adaptive streaming over HTTP in vehicular environments, in Proc. of the 4th Workshop on Mobile Video (MoVid), Chapel Hill, NC, USA, Feb. 2012, pp. 37-42.

[37] P. Juluri, V. Tamarapalli, and D. Medhi, SARA: Segment aware rate adaptation algorithm for dynamic adaptive streaming over HTTP, in Proc. of the IEEE Int. Conf. on Communication Workshop (ICCW), London, UK, Jun. 2015, pp. 1765-1770.

[38] 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 & Tutorials, 21-1 (2019) 562-585.

[39] K. Spiteri, R. Urgaonkar, and R. Sitaraman, BOLA: Near-optimal bitrate adaption for online videos, in Proc. of the Int. Conference on Computer Communications (INFOCOM), San Francisco, CA, USA, Apr. 2016.

[40] Ghent University, 4G/LTE Bandwidth Logs, available online: http://users.ugent.be/~jvdrhoof/dataset-4g, accessed: Sep. 2019.

[41] Blender Foundation webpage, available online: https://www.blender.org/foundation, accessed: Sep. 2019.

[42] Y. Shuai and T. Herfet, A buffer dynamic stabilizer for low-latency adaptive video streaming, in Proc. of the Int. Conference on Consumer Electronics, Berlin, Germany, Sep. 2016.

[43] Mobile Video Service Performance Study, HUAWEI White Paper, available online: http://www.ctiforum.com/uploadfile/2015/0701/20150701091255294.pdf, published: 2015, accessed: Sep. 2019.

[44] Github website, ITU-T Rec. P.1203 Implementation, available online: https://github.com/itu-p1203/itu-p1203, accessed: Sep. 2019.

[45] Github website, ITU-T Rec. P.1203 Codec Extension to VP9 and HEVC, available online: https://github.com/Telecommunication-Telemedia-Assessment/itu-p1203-codecextension, accessed: Sep. 2019.

[46] University of Waterloo webpage, The SSIMplus Index for Video Quality-of-Experience Assessment, available online: https://ece.uwaterloo.ca/~z70wang/research/ssimplus, published: Nov. 2014.

[47] C. Bampis, Measuring Video Quality with VMAF: Why you should care, AOMedia Research Symposium, San Francisco, Oct. 2019.

[48] 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, 29-1 (2019) 183-197.

[49] S. Tavakoli, S. Egger, M. Seufert, R. Schatz, K. Brunnström, and N. García, Perceptual quality of HTTP adaptive streaming strategies: cross-experimental analysis of multi-laboratory and crowdsourced subjective studies, IEEE Journal on Selected Areas in Communications, 34-8 (2016) 2141-2153.

[50] C. Moldovan, K. Hagn, C. Sieber, W. Kellerer, and T. Hoßfeld, Keep calm and don't switch: about the relationship between switches and quality in HAS, in Proc. of the Int. Teletraffic Congress (ITC), Genoa, Italy, Sep. 2017.

[51] GitHub website, Dashgen, Multimedia Communications Group, available online: https://github.com/comm-iteam/dashgen, accessed: Sep. 2019.