2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

QoE Models in the Wild: Comparing Video QoE Models Using a Crowdsourced Data Set

Anika Seufert*, Florian Wamser*, David Yarish†, Hunter Macdonald†, Tobias Hoßfeld*
* *University of Würzburg, Chair of Communication Networks*, Würzburg, Germany
{anika.seufert|florian.wamser|tobias.hossfeld}@uni-wuerzburg.de
† *Tutela Technologies, Ltd.*, Victoria, Canada
dyarish@tutelatechnologies.com, hmacdonald@tutela.com

***Abstract*—Crowdsourced measurements solve the problem of being able to assess the performance of a communication network from an end-user perspective, but the new characteristics of the data pose new challenges for QoE modeling. In contrast to existing laboratory or network measurements, this type of measurement at the end user device primarily involves taking a large number of short sample measurements, which, however, are rich in measured parameters, including many user-, application-, and device-related parameters. To test the applicability and to facilitate the integration of such data, we applied four QoE models from the literature to 290k worldwide video streaming measurements from a commercial data set from August to October 2020. In this work, we will therefore first describe the crowdsourcing video streaming data set to provide insights into the properties of video streaming KPIs in the real world. Second, we run four popular QoE models using this data set, compare the resulting QoE scores, and derive the impact of individual KPIs for each model. We show that the models assess the QoE at least differently, but sometimes with contradicting statements. Reading this paper, it becomes evident that more work and subjective studies, based on real-world data like the one we have shown, are needed to extend the current QoE models.**

***Index Terms*—Video Streaming, QoE, QoE Models, Crowdsourced Measurements, DASH**

I. Introduction

Crowdsourced measurements are the new trend when it comes to measuring user satisfaction on a large-scale [1,2]. In contrast to classical testing, here, the end users' devices are utilized to perform crowdsourced video streaming measurements (CVSMs) without the need of a subjective rating of the user. This measurement technique enables the collection of large amounts of data and allows network and service providers to estimate the current user experience. This is done by collecting key performance indicators (KPIs) during the playback of the video, which can later be converted to QoE scores using QoE models. Examples of CVSM providers include, for example, Tutela Technologies Ltd., who periodically conducts measurements in the background of several popular applications at over 300 million smartphones, or 5GMARK, which uses active tests on their own application for Android and iOS, which has been downloaded more than 1M times from Google Play Store.

When calculating QoE scores, it is important to use standardized QoE models to obtain meaningful and comparable results. The model P.1203 [3], which has been standardized by the ITU-T, would be suitable here, but it is restricted in its application. For example, it is only designed for videos having a duration between 1 min and 5 min with an initial delay of up to 10 s. As crowdsourced measurements are run at the end user's device in the cellular network, it is important to keep the energy consumption as well as the data usage as low as possible. Thus, crowdsourced measurement providers use shorter videos of about 30 s in their measurements. Furthermore, due to the fact that crowdsourced measurements are conducted in the wild, a wide range of KPI characteristics can occur, e.g., initial delays of more than 10 s. Hence, it is not clear if P.1203 is suitable for this kind of data and there is no model specially customized for the ever-increasing crowdsourced measurement market. In order to check if nevertheless verified statements about the QoE of end users can be made, we examine if they agree on the calculated QoE, i.e., if and which differences they show, using a large crowdsourced data set.

The contribution of this work is twofold. First, we give unprecedented insights into a large CVSM data set with more than 290k mobile crowdsourced measurements from around the world. Second, to check the applicability of existing QoE models, we compare four well-known QoE models with each other and then investigate the influence of the interaction of different KPIs on the individual models. We show that the models differ greatly, both in their distributions and in terms of individual scores and consideration of KPIs, so that no conclusion can be drawn about end-user satisfaction. We recommend extending existing models and conducting further subjective studies using the described KPI characteristics.

The remainder of this work is structured as follows. Section II provides background information and related works on crowdsourced measurements and QoE models by describing four popular QoE models from literature. In Section III, a CVSM data set is presented. Next, in Section IV, the models are compared based on their calculated scores of the presented crowdsourced data set. First, the scores per QoE model are directly compared and afterwards, the influence of specific video KPIs on the models is discussed. Finally, Section V concludes this work.

978-1-6654-3589-5/21/$31.00 ©2021 IEEE

II. Background and Related Work

For service and network providers, the QoE of their end users is the key factor for success. To avoid expensive subjective studies and to be able to collect real QoE values on a large scale, they increasingly use crowdsourced video streaming measurements (CVSM). The term crowdsourcing includes the participation of volunteers in an outsourced campaign. In the context of video measurements, they are defined as the measurement of video related key performance indicators (KPIs) by the crowd of mobile subscribers. There are different ways how to conduct crowdsourced video measurements. In [4,5], for example, CVSMs are conducted to collect crowd data on the smartphone of end users. Other commercial data providers publish crowdsourced video data performance statistics, for example Tutela, Ookla, or 5GMARK, where smartphones of end users are used as measurement device.

In order to draw conclusions from the collected KPIs to the QoE of end users, QoE models are used. In literature, various QoE models are proposed, which are summarized, e.g, in [6,7]. For a CVSM data set, only parametric models are suitable that operate on transport or application-layer parameters of the video streaming. However, most of the models were developed using only relatively small, limited data sets and have not been tested for their applicability to different large-scale use cases in mobile networks. The standardized P.1203 model, for example, is defined for videos between 1 min and 5 min, and therefore is not directly applicable for CVSM data with its typically short video sample measurements and wide range of KPI characteristics. For our further evaluation, we picked four models based on their applicability and popularity, as can be seen in Table I. The three of the most common models found in literature are the standardized *P.1203* [3], the *Petrangeli* [8], and the *MPC Model* [9]. In contrast to that, a simple model with low implementation complexity, which is nevertheless often used, is the *FTW Model* [10,11]. All four models comprise modules for short-term quality estimation for video streaming and predict a QoE score of an average user on a 5-point absolute category rating (ACR) scale.

FTW Model: Works like [12] state that the main influencing factors for their own subjective data set are mainly a combination of stalling and bitrate (correlation of 0.7264). The model of Hoßfeld *et al.* [10,11] focuses on stalling in the first hand and is only defined for non-adaptive streaming. Nevertheless, it is not complex, reflects the influence of one of the main influence factors for adaptive streaming, and can be easily applied to our large data set. It is used for illustration purposes here. It considers the length and any number of stalling events and is calculated as follows: $QoE_{FTW} = 3.5e^{-(0.15\psi+0.19)*\phi} + 1.50$ with ϕ is the number of occurred stalling events and ψ the average stalling event duration in seconds.

ITU P.1203 Model: The P.1203 model [3], which was standardized by ITU-T in 2017, is an adaptive streaming model, which takes not only stalling events into account, but also the audiovisual quality, e.g., bitrate and resolution. For this model, we used the implementation from [16]. *Mode 0* was used, as we only have access to the video's codec, target bitrate, coding and display resolution, frame rate, and segment duration. Furthermore, as our data set does not provide information about the audio quality, we set the audio quality module to the highest value as also recommended in the standard. To the best of our knowledge, only this model explicitly allows the estimation of QoE considering the particularities of mobile devices. Furthermore, P.1203 has been extensively validated with subjective studies. Since this model is the only standardized model, it is the only model for which the scope of application is clearly defined. However, these definitions indicate that it is not suitable for a CVSM data set, see, e.g., video duration of 1-5 min. Also the KPI characteristics are restricted. For example, only for initial delays up to 10 s, a maximum of 5 stalling events with a maximum event length of 15s, and a minimum video resolution of 240p. In the following, this model will be called $QoE_{P.1203}$ model.

TABLE I: Characteristics of the selected QoE models

Model	**# Citations**	**Stalling**	**Adaptation**	**Initial Delay**	**Picture Quality**	**Memory Effects**	**Complexity**
FTW [10,11], 2013/20,	174	✓					0
ITU-T P.1203 [3], 2017/20		✓	✓	✓	✓	✓	+++
MPC [9,13], 2015	1114	✓		✓	✓		+
Petrangeli [8,14,15], 2015	292	✓	✓		✓		+

Petrangeli Model: In the model of Petrangeli *et al.* [8,14,15], the QoE calculation consists of a video quality module and a buffering module. Here, the average used quality level $\overline{q}$ and its standard deviation $\hat{q}$ is used, normalized with respect to the highest available quality level $q_{max} = 6$. The buffering module F is calculated as follows: $F = \frac{7}{8}max(\frac{ln(\phi)}{6} + 1, 0) + \frac{1}{8}\left(\frac{min(\psi,15)}{15}\right)$, where ϕ is the number of occurred stalling events and ψ the average stalling event duration in seconds. Using both, the quality and the buffering module, the video QoE is thus calculated as $QoE_{Pm} = max\left(\frac{5.67\overline{q}}{q_{max}} - \frac{6.72\hat{q}}{q_{max}} + 0.17 - 4.95F, 0\right)$.

MPC Model: The last of the used models is a Model Predictive Control (MPC) model presented bei Yin *et al.* [9], which was further used by Mao *et al.* [13]. This model considers the used bitrate, number of bitrate changes, and stalling events and is defined as follows:
$QoE_{MPC} = \sum_{n=1}^{N} R_n - \mu \sum_{n=1}^{N} T_n - \sum_{n=1}^{N} |R_{n+1} - R_n|$ with $N = \textit{video length in seconds}$, $R_n = \textit{bitrate per second}$, $\mu = 4.3$, and $T_n = \textit{stalling time per second}$.

In [17], the authors conducted a parameter study to investigate the influence of well-known DASH KPIs on P.1203. However, as they did a parameter study, they did not analyze the combined effects of the QoE factors on the model, but considered single factors separately with synthetic input data. Furthermore, the compatibility of QoE models was investigated in [18]. Here, by comparing the calculated QoE scores of a popular web QoE model with the video streaming QoE calculated according to P.1203, the authors showed that even under the same network conditions, the models calculate

TABLE II: Overview of data set KPIs

	Mean	Std	Min	25%	50%	75%	Max
Initial delay [s]	2.31	4.69	0.05	0.94	1.29	1.93	116.52
# Stalling	0.15	0.47	0	0	0	0	6
Tot. stall. duration [s]	1.34	6.78	0.00	0.00	0.00	0.00	114.70
Most used bitr. [Mbps]	2.08	2.12	0.06	0.62	1.26	2.31	6.44
Mean bitr.[Mbps]	1.73	1.43	0.06	0.90	1.22	1.81	6.44
Most used res. [p]	861.49	290.05	144	720	1080	1080	1080
# Quality changes	0.96	0.89	0	0	1	1	25
# Quality degradations	0.23	0.54	0	0	0	0	13

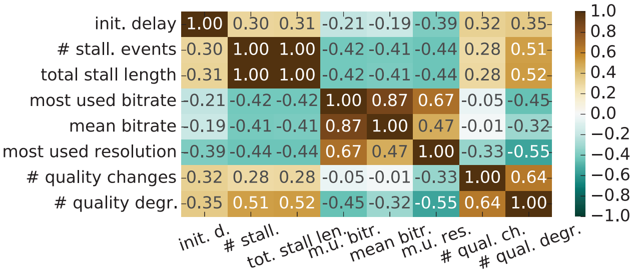


Fig. 1: Correlation of individual KPIs

different scores. Therefore, the question arises whether similar things can be seen when comparing different video streaming models with each other.

III. Crowdsourced Data Set

For our work, a commercial, large-scale data set from the independent crowdsourcing data company Tutela Technologies Ltd. is used. Tutela conducts DASH video tests all over the world to measure the video quality at end user devices.

The tests run in the background of mobile apps. To reduce the impact of the video test on a user's device or data usage, video tests are restricted to one per device per month and the video duration is set to 30 s. The tests are performed as follows: After randomly selecting the video platform (YouTube or Facebook), the video test is initiated at the end user's smartphone in the background of a mobile application. For this, the native video player in the Android or iOS operating systems is utilized, for Android devices Google's ExoPlayer, for iOS AvPlayer. Afterwards, the selected video is streamed using adaptive bitrate streaming. During the video playback, several application layer video streaming KPIs are monitored. In addition to device and network information in anonymous form, this includes the duration of the initial delay, the number of stalling events and its total duration, mean and most used bitrate, most used resolution as well as the number of quality changes and quality degradation numbers.

The underlying data set consists of 884,436 measurements for streaming video on cellular networks around the world. The data set was collected during three months from August to October 2020 from 205 different countries. Most measurements were conducted in the United States (14.23%), followed by the France (7.38%) and India (6.50%). After filtering out all incomplete measurements, e.g. measurement with interruptions due to handovers from the mobile network to WiFi or hardware issues during rendering of the video, 292.306 measurement remain.

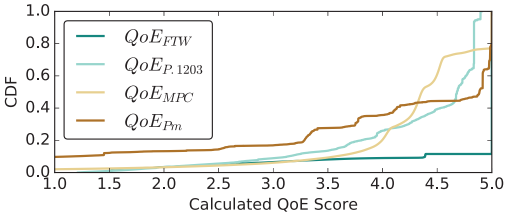


Fig. 2: Distribution of calculated QoE scores per model

The main performance parameters measured in the data set and their properties are listed in the Table II. Since the video measurements were performed under real world conditions on the end user's smartphone via the cellular network in an uncontrolled setting, all KPIs are included in the measurement data to varying degrees. For example, the data set includes measurements with zero as well as with up to six stalling events with a total length of up to 114.7 s. Nevertheless, no stalling occurred in most of the measurements. The used video quality ranged between 144p with 55 kbps up to 1080p with 6438 kbps and switched up to 25 times. Thus, it becomes clear that the classical KPI limits, such as those defined for P.1203, are not sufficient here, even though the videos are only half as long as the videos for which the model is defined.

While subjective QoE studies often focus only on one single performance indicator, in the wild, the interaction of the various KPIs with each other is of particular importance. Thus, Figure 1 show the coefficient of correlation of the most important KPIs. Here, higher negative correlations are depicted by darker green shading up to higher positive correlations are depicted by darker brown shading. Obvious strong correlations can be found between the number and total length of stalling events, between most used, average bitrate and most used resolution, as well as the number of quality changes and quality degradations. However, moderate correlations between other performance indicators can also be seen. For example, most used bitrate, mean bitrate, and most used resolution are negatively correlated to the number of stalling events as well as the total stalling event duration.

IV. Comparison of QoE Models

In order to draw conclusions about the QoE of the user based on this CVSM data set, models are required that calculate the QoE of an average user from the given KPIs. As explained above, however, there is yet no suitable model that formally corresponds to all the circumstances of such a data set. Therefore, we will use our crowdsourcing data set in the following to compare the resulting QoE scores of four models and to consider the influence and correlations of the KPIs measured worldwide on these models.

A. QoE Score Distributions per Model

To determine whether the QoE models are in consensus in terms of calculated QoE, the data set was used to calculate

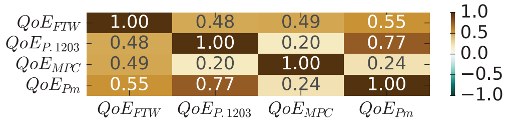


Fig. 3: Spearman rank-order correlations (SRCC) between QoE models

the QoE scores of each video using the four models. Figure 2 shows the distribution of the calculated QoE scores per model. The calculated QoE score is displayed on the x-axis, while the cumulative distribution function (CDF) is shown on the y-axis. As can be seen at first glance, the distributions of the models are different. The visible steps in the curves result from common combinations of the measured values contained in the data set, which are evaluated for different models to QoE clusters with the same value. While, for example, using QoE_{FTW} 88.49% of the measurements result in the top score of 5.00, for all other models the percentages are significantly lower, namely 20.00% for QoE_{Pm} and 22.91% for QoE_{MPC}. Note that the FTW model is not defined for adaptive streaming, but focuses solely on stalling length and number. It is reasonable cited and was included in the evaluation on the basis of the results of [12], which states that stalling is one of the most important influencing factors for QoE. Stalling percentage together with bitrate provides a Spearman's rank-order correlation coefficient (SRCC) of 0.7264 for the subjective SQoE-III database with simple linear regression. Using $QoE_{P.1203}$, not a single top score of 5.00 was calculated. The models also differ in their distribution of QoE scores. For example, QoE_{Pm} is the only model in which more than 10% of scores are bad (1.00). Only when considering the mean value per model, the values are relatively equal. Here, except for QoE_{FTW} having a mean of 4.75, the QoE score range between 4.23 and 4.29.

When looking at the absolute distance between the QoE scores for the individual videos, we can see differences. On average, the difference between the models is between 0.51 and 0.91. In addition, there are also videos where the calculated QoE score differs for all videos by 3.48 to 4.0. This means, that some KPI combinations result in a bad score of 1.00 for one model, but an excellent score of 5.00 for another. This disagreement between the models does not allow us to make any statements about the actual QoE of the end user, as all models claim to be validated in some way, but justifies further investigations.

The question arises whether the models, show broadly the same tendencies. For this reason, we evaluated the SRCC between the models in Figure 3. Relatively strong correlations can be seen for $QoE_{P.1203}$ and QoE_{Pm} (0.77) as well as QoE_{FTW} and QoE_{Pm} (0.55). In contrast to that, a very low correlation is calculated for QoE_{MPC} and $QoE_{P.1203}$ as well as QoE_{MPC} and QoE_{Pm} (0.20 and 0.24, respectively).

To analyze the relationship between the models in more detail, a direct comparison of the distributions of the calculated QoE scores per model are shown in Figure 4. The intensity of the color indicates where clusters occur, i.e., the darker the points, the more frequently they occur. In the optimal case, a linear relationship would result. No clear partial tendencies are visible, except for some weak correlations as in the combination of QoE_{FTW} and QoE_{MPC}, which is blurred by other values outside the diagonal, so that the SPCC results in 0.49. However, there is a strong accumulation for $QoE_{FTW} = 5$, whereas for QoE_{MPC} the corresponding scores scatter strongly. That is why the coefficient of correlation of these models is only moderate. Another example is the relationship between QoE_{FTW} and QoE_{Pm}. Here, the correlation matrix (Figure 3) showed a strong correlation of 0.55, but the correlation is only due to the high percentage of scores of 5.00.

To sum up, the premise here was: No model is more correct than the other if and only if the model has been subjectively validated in a statistically reliable manner. Nevertheless, we see differences due to the different nature of the models. No model can be rated better or worse, since we have no insight into the subjective QoE, which is typical for CVSMs. Insights can only give subjective studies as, for example, used in [3] or [12]. The fact of the different results, however, raises the question of the area of application of each of the models. We deduce from this the need for further subjective studies to focus the models themselves more closely and precisely with regard to their validated area of application. Furthermore, CVSMs in particular offer a wide range of application-related and device-related KPIs, which raises the question of whether other parameters would be helpful for the estimation.

B. Influence of KPIs on QoE Models

To provide an objective quantification of the differences of the models, we analyze the influence of the given crowdsourced video KPIs on the calculated scores. Figure 5 shows radar charts for displaying multivariate data, having one axis for each KPI showing the absolute correlation coefficient from 0 to 1 for each model. The higher the correlation coefficient is, the stronger the influence of this KPI on the respective model. A literature study led us to the KPIs shown on the axes, as these represent the commonly used input KPIs for the models in general.

Strong correlations can be seen between QoE_{FTW} and the number as well as the total length of stalling events, as they have an absolute correlation of 1.00. This is given directly by the definition of the model. All other KPIs for this model show weak correlation (correlation coefficient ≤ 0.5). The reason for this here is due to the fact that the individual KPIs correlate with each other, see Fig. 1.

Looking at $QoE_{P.1203}$, the most important KPIs cannot be identified as clearly as in the previous model. An absolute correlation of 0.63 to the number of quality changes is visible, which is in-line with with the results of [17]. Other KPIs, which show a moderate absolute correlation, are the number and total length of stalling events (0.47 and 0.48, respectively) as well as the most used resolution (0.48) and the number of quality changes (0.49). Although the bitrate is also included

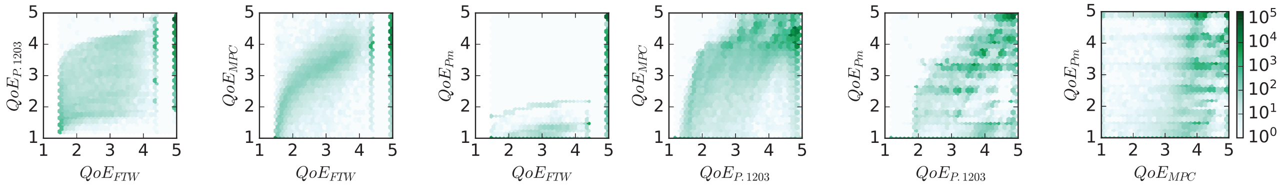
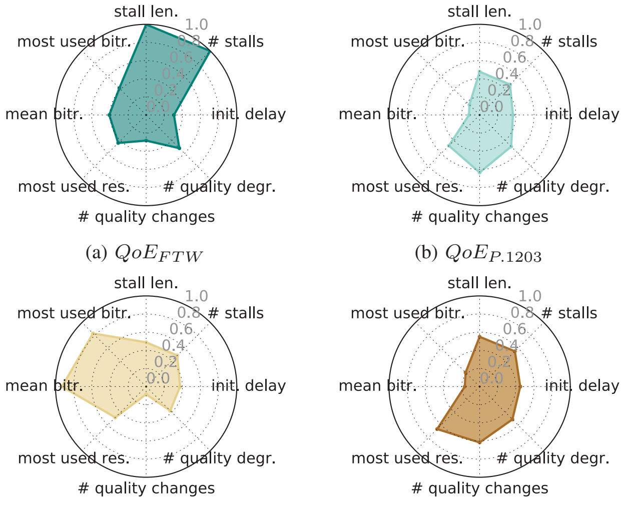


Fig. 4: Comparison of distribution of calculated QoE scores.



(a) QoE_{FTW} (b) $QoE_{P.1203}$

(c) QoE_{MPC} (d) QoE_{Pm}

Fig. 5: Influence of given video KPIs on model results by absolute correlation coefficients

in the calculation, it does not play a strong role in the model. Furthermore, the observation from [12] can also be applied here, where single KPIs only show moderate correlations to the subjective QoE, but combinations of parameters correlate with the QoE (regression model [12] with a combination of stalling percentage, bitrate, average bitrate switch magnitude shows SRCC of 0.7743 to subjective data). For further evaluations of the influence of KPIs in P.1203, we refer to [17].

With QoE_{MPC}, the mean and the most used bitrate play a key role in the calculation of the QoE score, showing an absolute SRCC of 0.83 (mean) and 0.93 (most used). For QoE_{Pm}, the model is not focused on one or two specific KPIs that thus show a very strong correlation, but on multiple KPIs showing strong to moderate correlations.

In Figure 6, the x-axes show the different KPIs while the y-axes show the four QoE models. Again, the intensity of the color indicates where clusters occur, i.e., the darker the points are the more frequently they occur.

Considering the total stalling time (third column), different results come up among the models. Here, QoE_{FTW} show a weak negative exponential and QoE_{MPC} a weak negative linear relationship, which results in a bad QoE to a different degree the longer the total stalling duration is measured. In comparison to that, for $QoE_{P.1203}$ and QoE_{Pm} the relationship is not clear. For $QoE_{P.1203}$, it is also possible that a relatively large stalling event duration leads to a fair QoE score. Looking at QoE_{Pm}, which showed a high correlation to the total stalling event length, a very drastic relationship is visible. Here, as long as the stalling event length is 0 s all QoE scores are possible, but as soon as stalling occur the probability of a bad score is very high.

TABLE III: Summary of the differences by comparing to the standardized P.1203 model

	Median	Key KPI	SRCC	MAPE	KS
$QoE_{P.1203}$	4.68	Adaptation	-	-	-
QOE_{FTW}	5.00	Stalling	0.48	0.15	0.88
QoE_{MPC}	4.78	Bitrate	0.20	0.15	0.31
QoE_{Pm}	4.89	Resolution/Adaptation	0.77	0.14	0.48

Another example where significant differences are visible is the mean bitrate (fifth column). QoE_{FTW} shows its inherent definition considering no other parameters than stalling length and number. A score of 5 is calculated regardless of the average bit rate. It is also noticeable that for a smaller mean bitrate, the scores vary greatly between bad and good scores. Some similar tendencies are visible for $QoE_{P.1203}$. In some cases, a bad mean bitrate leads to a good QoE score, whereas here a high mean bitrate does not necessarily lead to a very good QoE score, but fluctuates in the range from 3.5 to 5, which shows that other factors also have an influence. Using the QoE_{MPC} model, which previously showed a very high correlation to the mean bitrate, a more clear relationship can be seen. For QoE_{Pm}, on the other hand, the values are much more dispersed. Even for a high mean bitrate, QoE scores of 1 are calculated.

It was shown that the calculated QoE scores of the selected QoE models differ in their distribution as well as in the impact of each KPI on them. Table III summarizes our findings by first, listing the mean QoE scores per model as well as the KPI with the most influence on the model. Second, the differences between the models, using the example of the difference to the standardized P.1203, are highlighted. Here, SRCC, the Mean Absolute Percentage Error (MAPE), and the Kolmogorov Smirnov (with $p < 0.001$) values are shown. As all values indicate great differences of the models, no clear statement can be made about the subjective perceived user experience. For this reason, we recommend that network and service providers do not simply use any QoE model to calculate the QoE of CVSM data but continue to rely on standardized models. However, these standardized models, such as P.1203,

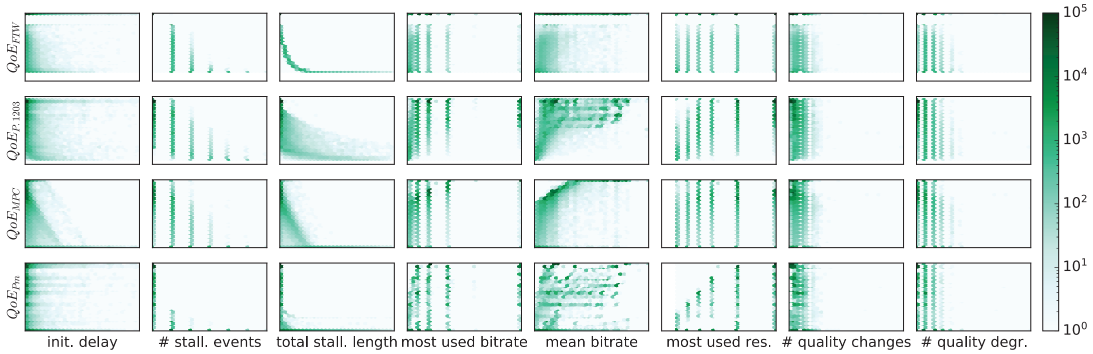


Fig. 6: Influence of different KPIs on QoE models

need to be extended to cover the use case of crowdsourced measurements with, for example, shorter video durations and a wider range of KPI characteristics. The previous described data set is ideal for this purpose, as it provides an overview of the real-world contexts and distributions of video streaming KPIs that can be used for subjective studies.

V. Conclusion

For video streaming service providers, the QoE of their end users is the key factor for success. Crowdsourced video streaming measurements lead to a new kind of data related to KPIs and volumes with certain characteristics like short video duration measured in uncontrolled environments. In literature, numerous models, especially also the standardized P1203 model, are presented which are based on different features, but, unfortunately, no QoE model exists which is formally defined for the area of CVSMs. Since CVSMs typically do not have any subjective ratings, i.e., no ground truth to compare, we compared the resulting QoE scores of four well-cited video streaming QoE models based on a large-scale crowdsourced data set and evaluated their influencing factors. Our results show that the calculated scores per model have significant differences, which are based on the different weighting of the KPIs. Since we cannot say anything about the validity of an single model due to the absence of subjective ratings, no definitive statement or recommendation can be made. However, the evaluations show that models used in the literature do show different results, which raises the need for new research on the applicability of QoE models. Therefore, caution is advised when comparing QoE values that are not known how they were calculated or that were obviously calculated using different models. In addition, we have shown how new crowdsourcing-based data sets look like and provide what characteristics they exhibit. This can be used to design QoE models for these particular measurement type of CVSMs.

References

[1] ITU-T, “Crowdsourcing approach for the assessment of end-to-end quality of service in fixed and mobile broadband networks,” ITU Telecomm. Standardization Sector, Standard ITU-T E.812:202005, 2020.

[2] T. Hoßfeld, S. Wunderer, A. Beyer, A. Hall, A. Schwind, C. Gassner, F. Guillemin, F. Wamser *et al.*, “White Paper on Crowdsourced Network and QoE Measurements–Definitions, Use Cases and Challenges,” 2020.

[3] ITU-T, “Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport,” ITU Telecommunication Standardization Sector, Standard ITU-T P.1203:201710, 2017.

[4] F. Wamser, M. Seufert, P. Casas, R. Irmer, P. Tran-Gia, and R. Schatz, “YoMoApp: A tool for Analyzing QoE of YouTube HTTP Adaptive Streaming in Mobile Networks,” in *European Conference on Networks and Communications (EuCNC)*. IEEE, 2015, pp. 239–243.

[5] A. Schwind, C. Midoglu, Ö. Alay, C. Griwodz, and F. Wamser, “Dissecting the Performance of YouTube Video Streaming in Mobile Networks,” *International Journal of Network Management*, p. e2058, 2019.

[6] N. Barman and M. G. Martini, “Qoe modeling for HTTP adaptive video streaming–a survey and open challenges,” *IEEE Access*, 2019.

[7] H. Zhang, F. Li, and Z. Yan, “Feature fusion quality assessment model for DASH video streaming,” *IET Image Processing*, 2020.

[8] 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 (TOMM)*, vol. 12, no. 2, pp. 1–24, 2015.

[9] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, “A control-theoretic approach for dynamic adaptive video streaming over HTTP,” in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 325–338.

[10] 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, pp. 264–301.

[11] T. Hoßfeld, P. E. Heegaard, M. Varela, L. Skorin-Kapov, and M. Fiedler, “From QoS Distributions to QoE Distributions: a System’s Perspective,” *arXiv preprint arXiv:2003.12742*, 2020.

[12] 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.

[13] H. Mao, R. Netravali, and M. Alizadeh, “Neural adaptive video streaming with pensieve,” in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, 2017, pp. 197–210.

[14] M. Claeys, S. Latré, J. Famaey, T. Wu, W. Van Leekwijck, and F. De Turck, “Design and optimisation of a (FA) Q-learning-based HTTP adaptive streaming client,” *Connection Science*, pp. 25–43, 2014.

[15] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, “Model for estimating QoE of video delivered using HTTP adaptive streaming,” in *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*. IEEE, 2013, pp. 1288–1293.

[16] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia *et al.*, “HTTP adaptive streaming QoE estimation with ITU-T rec. P. 1203: open databases and software,” in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 466–471.

[17] M. Seufert, N. Wehner, and P. Casas, “Studying the impact of HAS QoE factors on the standardized QoE model P. 1203,” in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2018, pp. 1636–1641.

[18] M. Seufert, N. Wehner, V. Wieser, P. Casas, and G. Capdehourat, “Mind the (QoE) Gap: On the Incompatibility of Web and Video QoE Models in the Wild,” in *2020 16th International Conference on Network and Service Management (CNSM)*. IEEE, 2020, pp. 1–5.