On user-centric analysis and prediction of QoE for video streaming using empirical measurements

Maria Plakia, Michalis Katsarakis, Paulos Charonyktakis, and Maria Papadopouli
Department of Computer Science, University of Crete
Institute of Computer Science, Foundation for Research and Technology-Hellas

Ioannis Markopoulos
Forthnet S.A.

***Abstract*—Assessing the impact of different network conditions on user experience is important for improving the telecommunication services. We have developed a modular framework that includes monitoring and data collection tools and algorithms for user-centric analysis and prediction of the QoE in video streaming. The MLQoE employs several machine learning (ML) algorithms and tunes their hyper-parameters. It dynamically selects the ML algorithm that exhibits the best performance and its parameters automatically based on the input (e.g., network and systems metrics). We applied the MLQoE for predicting the QoE of the video streaming service in the context of two field studies, one performed in the production environment of a large telecom operator and the other at our Institute. The analysis indicated the parameters with the dominant impact on the perceived QoE and revealed that the QoE vary across users. This motivates the use of customized adaptation mechanisms in video streaming under network performance degradation. The MLQoE results in fairly accurate predictions e.g., a median error in predicting the QoE of 0.0991 and 0.5517 in the first (second) field study, respectively, on the MOS scale.**

I. Introduction

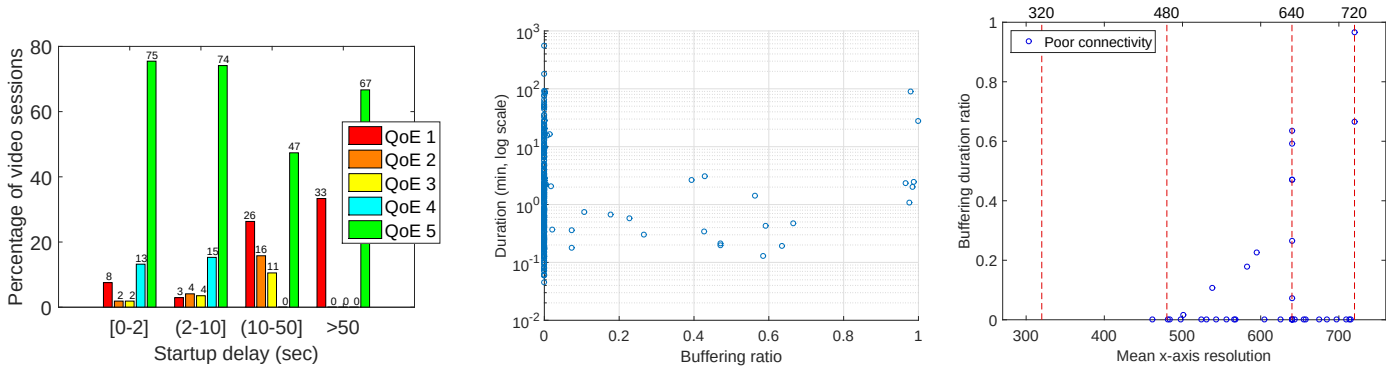
The impact of the network performance on the quality of experience (QoE) for various services is not understood in depth. The QoE can be defined as "the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the persons evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the persons context, personality and current state" [1]. This definition reflects some of the user-centric and contextual aspects of QoE. In general, depending on the type of service and the context, the QoE can be affected by various techno-socio-economic-cultural-psychological parameters, e.g., by the user preferences with respect to QoE and price, willingness-to-pay, and intrinsic indicators towards a service provider (e.g., band name, perceived value, reliability), its content (e.g., richness, diversity, searching mechanisms), and even integration with other popular services (e.g., social networking applications). It may be difficult to dynamically capture these aspects and assess to which extend they affect the QoE of a service, especially in a non-intrusive manner. Thus, the design of the appropriate metrics and methodologies to monitor the infrastructure (e.g., network, system, and context), collect the appropriate data, and model the QoE can be challenging.

Our community has been assessing the impact of the network on the user experience, which is critical for improving the telecommunication services. A diagnostic tool that indicates whether users perceive the deterioration of the network performance can be very useful. When users do not perceive performance degradation, an adaptation could be avoided. Moreover for churn prevention, cost reduction, increasing revenue, rolling out new services and differentiating their existing ones, the knowledge about the user engagement and satisfaction is important in order to create competitive advantages within the Internet market. Characterizing the QoE for VoIP, video streaming, and web browsing, has been at the epicenter of various activities. For example, the prediction of QoE for video can be performed by applying mathematical models based on QoS parameters [2]–[4], signal processing techniques [5] or data-mining algorithms [6]–[11]. The majority of such efforts aim to characterize the user experience, analyzing measurements in an aggregate manner.

We recently developed a modular algorithmic framework for user-centric QoE prediction, MLQoE [12]. This framework employs multiple machine learning (ML) algorithms, namely, Artificial Neural Networks (ANN), Support Vector Regression (SVR) machines, Decision Trees (DTs), and Gaussian Naive Bayes (GNB) classifiers, and tunes their hyper-parameters. It selects the ML algorithm that exhibits the best performance and its parameters automatically, given the input. The input involves network and systems metrics based on empirical measurements as well as subjective opinion scores collected from users. In an earlier work, we analyzed the performance of this framework on VoIP and video traces from the LIVE Mobile VQA database, which consists of a number of reference and distorted videos. The distorted videos have been created by varying the compression rate, rate-adaptation, number of frame-freeze, and packet-loss. 18 subjects assessed the quality of some of the distorted videos. This work further extends our earlier research in several directions: we developed the QoE tracker, a monitoring and data collection system. In the context of a video streaming service on mobile devices provided by a large telecom operator in Greece, in its production environment, we performed the first field study. Volunteers employed the QoE tracker and evaluated the perceived QoE of the video streaming service. We analyzed the collected data to understand the impact of various parameters, such as startup

* This work is supported partially by the General Secretariat for Research and Technology in Greece with a Research Excellence, Investigator-driven grant, 2012-2015 and by Forthnet S.A (PI Maria Papadopouli). Contact author: Maria Papadopouli (mgp@ics.forth.gr).

978-1-5090-0354-9/16/$31.00



(a) (b) (c)

Fig. 1. (a) Histogram of the QoE distribution for different startup delays, (b) the duration of the session as a function of the buffering ratio, and (c) buffering ratio as a function of the weighted mean video resolution for the sessions terminated by poor connectivity.

delay, rebuffering events, packet losses, and changes of resolution, on the QoE. This work also focused on the user-centric aspects in QoE and sensitivity of users to different type of impairments. We then applied the MLQoE to predict the QoE score and compared its performance with the Weber-Fechner Law (WFL) [2]. This field study took place in a dynamic "open" relatively unrestricted and heterogeneous environment, which imposed several challenges in the analysis. To validate the outcome of the analysis and further extend it, we performed a second more controlled field study in our Institute. The paper evaluates the performance of the MLQoE prediction using the data collected in the second field study and highlights the main results. The paper is structured as follows: Section 2 describes the field studies. Section 3 focusses on the analysis of the field studies. Section 4 presents the proposed methodology, while Section 5 discusses the performance analysis. Section 6 overviews the related work. Section 7 summarizes our conclusions.

II. QoE tracker

A major Greek telecom operator has been providing a VoD, LiveTV, TSTV, and TVoD video streaming service. The video streaming service employs the HTTP Adaptive Streaming (HAS) technique [13]. Also, the HAS technology uses the TCP protocol for reliable video transmission. In a joint project, we developed the QoE tracker, a monitoring system that collects network and systems measurements (objective measurements) as well as feedback from users (subjective measurements). The QoE tracker follows the client-server architecture. It runs on the smartphone of the user (*client*), monitors the network in the background, and parses the log messages generated by the video streaming client, when the user performs certain actions. At the end of a video viewing session (from now on called *session*), the user rates the session by providing an opinion score (on the MOS scale). The collected measurements "capture" various events, such as resolution changes, buffering events, and user actions with respect to video viewing. The QoE server (*server* from now on) is running on a Linux virtual machine and collects, stores and analyzes the objective and subjective data uploaded by the clients. The client consists of the monitor, GUI, performance estimator, database, and the back-end interface. The monitor is composed of three sub-modules, namely, the logcat parser, active prober, and localization. The logcat parser parses periodically the log messages of the video streaming client, recognizes various user actions and other events that may occur during the session, and keeps track of the state of the video player. When a video session start (end) is identified, the active prober is launched (terminated), respectively. During its activation, the active prober communicates with the active prober module of the server, for the initiation of network measurements through the iperf tool. The localization sub-module determines the geographical location of the device during a video session. The video streaming client uses only the wireless network (i.e., WiFi). Similarly the communication of the QoE system takes place via wireless network.

For each video session, the following features were collected: the *service type, startup delay*, *the ratio of the startup delay over the session duration, session duration, QoE score, number of buffering events* (and statistics about the duration of them, such as total, min, max, mean and standard deviation), *the mean weighted resolution* and *the ratio of the weighted mean video resolution over the size of the display of the user device, the number of switches of the video resolution* (and statistics based on them, e.g., min, max, mean and standard deviation), *packet loss, jitter,* and *signal strength*. The user activity is characterized by the duration of the pause, seek, and off-screen events. The same statistics are also computed for the *last* 15, 30, and 60 sec of the session. The termination type which indicates whether the session was terminated due to poor connectivity or normally by the user is also obtained.

III. Analysis

The first field study took place in the context of the video streaming service provided by a large Greek telecom operator. During this field study (that lasted 56 days), 20 volunteers, customers of the service, participated by viewing videos, uploading at least one *labelled* video session. We consider as labelled session a session that has been rated with a QoE score by the user. The devices of the participants vary in terms of their manufacturer, model, display size, and Android version. The collected dataset includes 293 stationary sessions and five

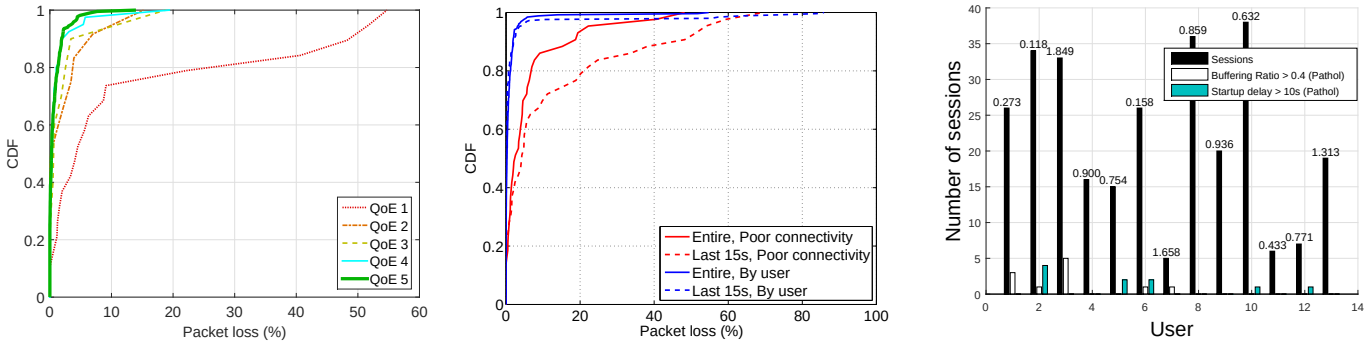


Fig. 2. Using the dataset from the first field study. Cumulative Distribution Function (CDF) of the (a) packet loss for sessions rated with different QoE scores, (b) packet loss for sessions with different termination types as well as packet losses considering only the last 15 sec of the sessions, and (c) the mean absolute error of the MLQoE per user, indicated at the top of each column, considering all his/her sessions (left column), the sessions with high buffering ratio and QoE score of 5 (middle), and the sessions with high startup delay and QoE score of 5 (right column).

wireless handover ones. The handover sessions exhibit higher packet loss and jitter, and stronger signal strength than the stationary ones. We first analyzed the impact of the startup delay on QoE. Users perceive the degradation (reflected by the low QoE scores) when the startup delay is 10 sec or more (Fig. 1 (a)). Prior related research (e.g., by Krishnan *et al.* [14]) reported that a startup delay beyond 2 sec causes viewers to abandon the video. The dataset in [14] contains measurements obtained from wired, wireless, and cellular connections. We speculate that the smartphone wireless network users of our study are perhaps more tolerant in the startup delay than users with fixed devices using a larger bandwidth connection. The higher the buffering ratio, the smaller the duration of the session (Fig. 1 (b)). This trend has been also observed in the related work (e.g., [14], [15]). Interestingly, there were several sessions of high resolution that terminated with poor connectivity and exhibit high buffering duration ratio and low QoE scores (Fig. 1 (c)). Potentially, by lowering the resolution, the buffering ratio could be reduced resulting in an improved QoE. As expected, sessions that experience worse network conditions (e.g., in terms of jitter, packet loss and RSSI), high startup delay, and buffering ratio are rated with lower QoE. The larger the packet loss, the lower the QoE score (Fig. 2 (a)). These are also more likely to terminate with a poor connectivity status (Fig. 2 (b)). Poor network performance during the last 15 sec of the session may result in termination due to poor connectivity. Specifically, sessions that have been terminated with poor connectivity exhibit higher packet losses, jitter, and buffering ratios *during their last 15 seconds* than the entire sessions terminated with poor connectivity or the sessions terminated normally by the user (Fig. 2 (b)).

There are sessions with high buffering ratio that lasted more than 10 min and were rated with a score of 5. Moreover, there were sessions of high QoE scores that were terminated with poor connectivity status or had a startup delay of 50 sec or more (Figs. 2 (a) and (c)). That is, even though users experienced a degraded performance, they still rated these sessions with high QoE scores. As mentioned earlier, the first field study took place in a dynamic heterogeneous and relatively unrestricted environment. We speculate that depending on the context and content of these sessions, the expectations and tolerance of the users may vary. However, this motivated us to perform a second (more controlled) field study in our Institute. For the second study, we produced a number of synthetic video sequences that correspond to a wide range of network conditions. The effect of the network conditions (such as packet loss, jitter, RSSI) on the systems parameters, such as startup delay, buffering ratio, and resolution, depends on the specific video codec and application. Moreover the systems parameters affect directly the user perceived QoE. We generated scenarios of different types of impairment by varying these parameters and created (playback) videos that "manifest" these impairments. We used four different reference videos, each corresponding to high quality (i.e., did not exhibit any type of impairment) and displaying a different scene. Each scene has a total duration of 20 sec, while each video consists of 4 chunks with a duration of 5 sec. Each playback video was parametrized based on the startup delay, number of buffering events, ratio of buffering duration, times when buffering events occur, duration of each buffering event, video resolutions for each chunk, and aggregate resolution of the video. The startup delay and the buffering ratio have been modelled according to the Bounded Pareto distribution. The parameters of the distributions were estimated based on the empirical measurements of the first field study. During a video, up to three buffering events may occur, one after each chunk. The resolution may remain fixed or vary during the video. For the second study, fifty video sequences were produced. 20 participants, volunteers, mostly graduate students in our Institute, assessed the quality of these sessions using the MOS scale. Due to the large number of videos to be assessed, each participant viewed the videos during two viewing phases that took place in different days. The subjects viewed and assessed the videos using an Android application implemented on a Nexus 5. To obtain demographic information (e.g., age, sex, frequency of use of mobile applications, video streaming services, audiovisual tests) about the volunteers, each subject had to first answer a short questionnaire. Before viewing the

videos, the subject had to read and follow the instructions that appeared on the screen. After that, the training video sequences appeared in order to familiarize the subject with the various types of audiovisual quality degradation. Then, the subject viewed each video sample and indicated his/her opinion score about its QoE via the Android application.

We also aimed to further explore the subjectivity of the assessments and sensitivity of users to different types of impairment (e.g., large startup delay, number of rebuffering events, low resolution). We considered three types of prominent impairments, namely, the large startup delay, number of rebuffering events, and low resolution, and created three homogeneous sets with respect to these impairments. Specifically, the set with the prominent startup delay includes *only* the video sessions of high startup delay, excluding the video sessions of large number of rebuffering events or low resolution. We also created the other two video sets in a similar manner. We then analyzed how users rate the QoE of these videos. Indeed it appears that depending on the type of impairment, some users are more tolerant or strict than others. We define that a user assessed a session in a *lenient (strict)* manner when his/her score belongs to the 90-th (10-th) percentile of the total scores for this video provided by all 20 users of the field study, respectively. A user is labelled as lenient (strict) when 50% or more of his/her sessions are rated in a lenient (strict) manner, respectively. Some users are persistently labelled as lenient (e.g., users 5 and 6) or strict (e.g., users 7 and 19) across all the three types of impairment. Moreover, some users (e.g., users 1 and 2) are more tolerant to some types of impairment (e.g., high startup delay and low resolution) but sensitive to others (e.g., buffering events). To evaluate if the difference of the scores of users for the various types of impairment is statistically significant, we applied the Student's T-test, on their QoE scores. For the persistently lenient and strict users, the QoE scores among the various types of impairment are not statistically significant different, while for users that are tolerant to only some types of impairment the QoE scores are statistically significantly different.

IV. MLQoE Algorithm

The MLQoE employs supervised regression, in which the predictors are metrics, e.g., based on jitter, packet loss, rebuffering, startup delay, resolution, and the predicted outcome is the QoE score. The predictors are determined based on the specific service, size of the collected data, characteristics of the testbed and measurement study. The MLQoE consists of several modules, including the normalization, feature selection, training multiple regressors, the selection of the best ML model and the estimation of its performance. It employs a set of ML algorithms, which can be easily extended. The MLQoE has two main phases, namely, the model selection and performance estimation. The model selection takes as input the training set of the performance estimation loop, cross-validates it, and reports the best model. The performance estimation obtains as input the dataset, partitions it into folds, estimates the performance of the best model (that the model selection outputs) in each fold and reports (as output) the mean error for the dataset. The performance metric is the *absolute difference* of the predicted QoE score compared to the actual score provided by the user (which serves as the "ground truth").

To address the high dimensionality of the data (i.e., reduce the number of metrics that have to be measured), the MLQoE employs the Max-Min Parents and Children (MMPC), a causal-based and Bayesian Network-based feature selection algorithm. The MMPC identifies the parameters that have a dominant impact on QoE. In the model selection phase, the MMPC selects its hyper-parameters, namely, the maximum size of conditioning set k and the statistical level for accepting probabilistic dependence a. Unfortunately, estimating the performance of multiple models on the same test set leads to overestimation of the performance of the best performing model. To provide a conservative estimation, while at the same time avoid underfitting, the MLQoE employs the Nested Cross-Validation (nested CV) protocol [16]. The data normalization and feature selection is executed inside the nested CV, them participating in the model selection procedure.

V. Evaluation of the MLQoE prediction

To evaluate the prediction accuracy of the MLQoE in video streaming service, for the first field study, we used the collected datasets and applied the MLQoE in an aggregate and a user-centric manner. In this dataset, only 13 out of 20 users have rated five sessions or more. We consider only these users for the performance analysis of the MLQoE. The aggregate approach considers all the video sessions for all users. Due to the small number of samples for some users in the user-centric approach (in the first dataset), we used the leave-one-out nested CV (LOOCV) with random partitioning to folds. For the aggregate approach, a 10-fold nested CV with random partitioning to folds was applied. For the second dataset (i.e., collected in the second field study), in the user-centric approach, we used a 10-fold nested CV, for the evaluation of the prediction, considering all 20 users (since each user has assessed 50 video sessions).

Apart from the original datasets, a normalized version is also maintained. The normalization is performed to handle the variability across the metrics. It transforms the values of each metric to fit a normal distribution of zero mean and unit variance. Each ML algorithm has a number of tuning parameters [12]. At the model selection process, all the combinations of the different parameters are tested. [1]

Parameter Impact In the aggregate MLQoE, in the context of the first field study, the MMPC indicates that the parameters with dominant impact on the QoE are the termination type of the session, the buffering events frequency, the weighted mean video resolution ratio, and the packet loss. Considering the 13 subjects of the first field study, the MLQoE reported the termination type as a dominant factor for 10 subjects, the

[1]The MMPC algorithm is tested with $k = 0, 1, 2, 3$ and $a = 0.01, 0.05, 0.1, 1$ (a value $a = 1$ corresponds to selecting all variables without feature selection). Each dataset is employed to train the ML algorithms. The LIBSVM library Version 3.14 was used for the implementation of the SVR algorithm; the hyper-parameters were chosen as follows: for the Gaussian, linear, and polynomial kernels were used with the default values, the cost C was selected among the values $\{0.01, 0.1, 1, 10, 100\}$, and the insensitivity parameter e within values $\{0.05, 0.1, 0.25, 0.5, 1\}$. The ANN was implemented with one hidden layer. The number of nodes for the hidden layer varied from 8 up to 11 and 2 up to 5 [17] for the first and the second field study, respectively. The CART implementation have been used for the DTs and we tested the following values of the pruning level alpha $a = \{0.1, 0.01, 0.05\}$.

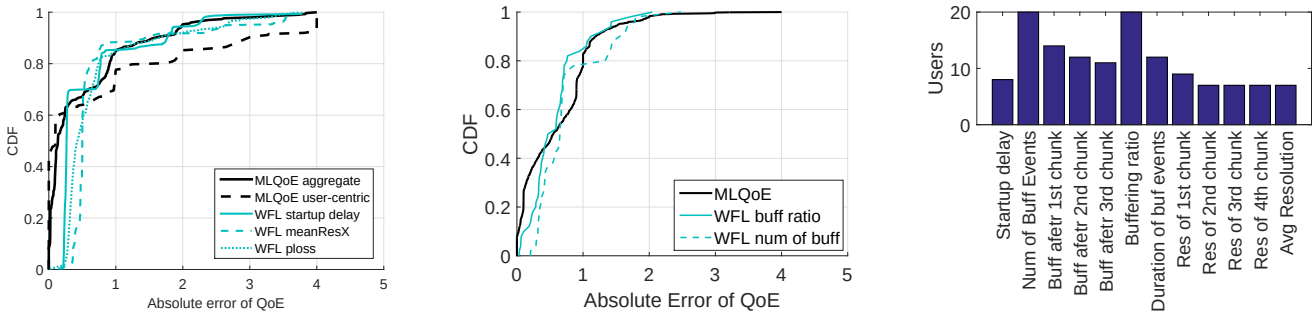


Fig. 3. The absolute error of the MLQoE and WFL for the first and second field study, (a) and (b), respectively, and (c) the features that have a dominant impact on QoE in the second field study.

mean jitter for 6 subjects, the startup delay and its ratio over the entire video duration for 5 subjects, the packet loss for 4 subjects, and the weighted mean video resolution and its ratio of the weighted mean video resolution over the display for 4 subjects. The diversity in the dominant parameters between the aggregate case and the user-centric one is due to the use of different datasets. The aggregate approach employs all the video sessions for all subjects, while in the user-centric approach, each subject views a different set of videos sessions, under different network conditions and context. In the second field study, the number of buffering events and buffering ratio consistently are the parameters with the most prominent impact on the QoE across all subjects (Fig. 3 (c)). The MLQoE captures the individual user preferences, as different dominant parameters are reporting for the different subjects.

Accuracy of the prediction For the first field study, the user-centric MLQoE can predict the QoE in a fairly accurate manner with a median and mean absolute error (MAE) of 0.0991 and 0.7716, respectively. The aggregate MLQoE reports a median and mean absolute error of 0.1392 and 0.5185 (Fig. 3 (a)). The better mean performance of the aggregate MLQoE compared to the user-centric approach is due to the specific limitations of this field study (as discussed also in Section III). For example, only ten users have rated more than 15 sessions. Moreover, there were users that provided only two or three different scores. These characteristics impact the training of the data mining algorithms and result in large prediction errors. For the second dataset, the user-centric MLQoE can predict the QoE score with a median and mean absolute error of 0.5517 and 0.6133, respectively (Fig. 3 (b)). The ML algorithm that exhibits the best performance may vary across users, so it is important to test various ML algorithms and a range of their hyper-parameters. [2]

We compared the MLQoE to the WFL, a state-of-the-art QoE model. The WFL reflects the relation of QoE and QoS using logarithmic regression [2]. The performance of the WFL was evaluated using a 10-fold cross-validation and one parameter as input (the packet loss or the mean resolution). The aggregate MLQoE exhibits a statistically significant better performance than the WFL in terms of mean and median prediction (Fig. 3 (a)), while the user-centric MLQoE outperforms WFL in terms of median prediction error. Although the WFL does not capture the interplay and impact of the multiple factors (e.g., mean packet loss, mean video resolution), it still has a reasonably good performance.

[2]For example, in the second field study, the best performing algorithm was the GNB, the SVR with Gaussian kernel, and the SVR with Linear kernel, for 7 (MAE 0.5762), 7 (MAE 0.6371), and 6 (MAE 0.6790) users, respectively.

VI. Related Work

The prediction of QoE for video can be performed by applying mathematical models based on QoS parameters (e.g., WFL and IQX [2]), full-reference algorithms (e.g., VQM [5]). For example, Hossfeld *et al.* [3] proposed a QoE model based on WFL for YouTube. Different types of relations between the QoS (network-level traffic characteristics) and QoE (e.g., linear, logarithmic, exponential and power) applied in [4] and shown the relationship between them. There are studies [14], [18] that used statistical tests (e.g., Pearson, Kendall) to evaluate the QoE based on user engagement, abandonment rate, and frequency of visits. Hands and Wilkins [19] examined the quality and acceptability for video streaming under different network conditions and showed that the burst size (number of consecutive dropped packets) has a considerable impact on QoE and acceptability. In [20], [21] they built applications that collect QoS (such as player state, statistics of buffering events, and video quality level) in the context of video streaming and web browsing services, parameters that impact the perceived QoE. The evaluation of acceptance, satisfaction, entertainment, and information recognition in different contexts (e.g., train station, bus) using ANOVA, Pearson correlation, Spearman, and Chi-square was the focus of [22]. The role of the context on QoE for various streaming services has been highlighted in several studies (e.g., [23]). Xue and Chen [24] evaluated the influence of contextual factors, such as display size, viewing distance, ambient luminance and user movement on subjective perceived quality. The context and the repeatability of the experiments was also analyzed in [25]. In the context of video streaming and telepresence, Wu *et al.* [26] characterized the QoS based on interactivity, vividness and consistency and the QoE using as metrics the concentration, enjoyment, telepresence, perceived usefulness, and perceived easiness of use and applied Pearsons correlation to map the QoS to QoE. Other studies use ML algorithms with

hold-out estimation [6], [7] or with cross-validation [8]–[10] and try to estimate the QoE. Simple regression models have been also used in order to characterize the user satisfaction [15], [27]. In general, the ground-truth for the QoE has been formed based on either the explicit opinion scores reported by users (e.g., in the context of audiovisual tests or at the end of their service via a GUI) or based on measurements collected using physiological metrics [28], [29]. Note that all the aforementioned models estimate the QoE for an average user in contrast to MLQoE that can be employed to capture also the individual user preferences. The closest paper in our work [11] builds the aggregated model training two different algorithms (DTs and ANN) using 10-fold cross validation for DTs and hold-out estimation for ANN. So the estimated performance is overestimated. They choose the best performed model and train it in a user-centric manner. The models have been trained using all the parameters, unlike our work that performs feature selection for dimensionality reduction.

VII. Conclusion and Future Work

The startup delay and buffering ratio affect the QoE. Sessions with startup delay higher than 10 sec obtain lower QoE scores, while sessions with buffering ratio, have typically a smaller duration. Sessions with poor network performance during the last 15 sec are likely to be terminate with poor connectivity. In several sessions, we observed that a rate adaptation could reduce the buffering ratio and improve the QoE. In the first field study, the parameters with a dominant impact on the QoE are the frequency of buffering events, weighted mean video resolution ratio, termination type, and packet loss (considering the aggregate prediction model). The median error in the QoE prediction is less than 0.1. In the second (more controlled) field study, the number of buffering events and buffering ratio are the parameters of prominent impact on QoE for each user. The sensitivity of users to the different types of impairment varies across users. Moreover we observed the presence of lenient and strict users (in terms of their QoE assessments).

The performance of these two field studies enabled us to reflect about the tradeoffs between small-scale studies with homogeneous settings in non-controlled environments and larger-scale (potentially crowd-sensing/sourcing participatory) studies that can reach more people, representing a more realistic set of conditions but with several unknown, difficult to control, exogenous parameters and heterogeneous settings. Obtaining reliable measurements in such crowd-sourcing non-controlled field studies can be challenging. In general, it is difficult to obtain the "ground truth" about the QoE. The above tradeoffs also highlight the tension between the subjectivity and reliability in the collected measurements.

Through the QoE tracker and proposed algorithms, the provider can learn more about its customers (e.g., their traffic, usage pattern, end-to-end network performance, QoE profile), infrastructure and service performance. This can enable the provider to improve the adaptation mechanisms, provide better customer service, assess its agreements with infrastructure/network providers, and potentially perform better pricing.

References

[1] S. Möller and A. Raake, *Quality of Experience*. Springer, 2014.

[2] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The logarithmic nature of qoe and the role of the weber-fechner law in qoe assessment," in *ICC*, 2010.

[3] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: between the devil and the deep blue sea," in *QoMEX*, 2012.

[4] J. Shaikh, M. Fiedler, and D. Collange, "Quality of experience from user and network perspectives," *annals of telecommunications*, 2010.

[5] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *Broadcasting, IEEE Trans. on*, 2004.

[6] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez, "Optimized online learning for qoe prediction," in *Proc. of BNAIC*, 2009.

[7] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez, "Predicting quality of experience in multimedia streaming," in *MoMM*, 2009.

[8] V. Menkovski, G. Exarchakos, and A. Liotta, "Online qoe prediction," in *QoMEX*. IEEE, 2010.

[9] D. Joumblatt, J. Chandrashekar, B. Kveton, N. Taft, and R. Teixeira, "Predicting user dissatisfaction with internet application performance at end-hosts," in *INFOCOM, 2013*.

[10] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Understanding the impact of network dynamics on mobile video user engagement," in *SIGMETRICS Performance Evaluation Review*, 2014.

[11] Y. Chen, Q. Chen, F. Zhang, Q. Zhang, K. Wu, R. Huang, and L. Zhou, "Understanding viewer engagement of video service in wi-fi network," *Computer Networks*, 2015.

[12] P. Charonyktakis, M. Plakia, I. Tsamardinos, and M. Papadopouli, "On user-centric modular qoe prediction for voip based on machine-learning algorithms," *IEEE Trans. on Mobile Computing*, 2015.

[13] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia, "A survey on quality of experience of http adaptive streaming," *Communications Surveys & Tutorials, IEEE*, 2015.

[14] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs," *IEEE/ACM TON*, 2013.

[15] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," *SIGCOMM Computer Communication Review*, 2011.

[16] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, 2005.

[17] T. M. Mitchell, "Artificial neural networks," *Machine learning*, 1997.

[18] M. T. Diallo, F. Fieau, and J.-B. Hennequin, "Impacts of video quality of experience on user engagement in a live event," in *ICMEW*, 2014.

[19] D. Hands and M. Wilkins, "A study of the impact of network loss and burst size on video streaming quality and acceptability," in *IDMS*. Springer, 1999.

[20] Q. A. Chen, H. Luo, S. Rosen, Z. M. Mao, K. Iyer, J. Hui, K. Sontineni, and K. Lau, "Qoe doctor: Diagnosing mobile app qoe with automated ui control and cross-layer analysis," in *IMC Conference*, 2014.

[21] 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 *Networks and Communications,*, 2015.

[22] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?" in *MobileHCI*. ACM, 2008.

[23] J. Hecht, "All smart, no phone," *Spectrum, IEEE*, 2014.

[24] J. Xue and C. W. Chen, "A study on perception of mobile video with surrounding contextual influences," in *IEEE QoMEX 2012*.

[25] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE J-STSP*, 2012.

[26] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, "Quality of experience in distributed interactive multimedia environments: toward a theoretical framework." ACM, 2009.

[27] K.-T. Chen, C.-C. Tu, and W.-C. Xiao, "Oneclick: A framework for measuring network quality of experience," in *INFOCOM 2009, IEEE*.

[28] G. M. Wilson and M. A. Sasse, "Do users always know whats good for them? utilising physiological responses to assess media quality," in *People and Computers XIVUsability or Else!* Springer, 2000.

[29] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & Information Technology*, vol. 25, 2006.