Improving QoE Prediction in Mobile Video through Machine Learning

Pedro Casas
AIT Austrian Institute of Technology
pedro.casas@ait.ac.at

Sarah Wassermann
Université de Liège
sarah.wassermann@student.ulg.ac.be

***Abstract*—Despite the massive adoption of HTTP adaptive streaming technology, buffering is still the most harmful event for QoE in video streaming. Previous studies have shown that buffering is not only detrimental for the overall user experience, but is also highly correlated to viewer engagement. The occurrence of buffering is particularly critical in cellular networks and mobile video deployments, as network conditions are less stable and network resources more limited. In this context, monitoring and properly predicting the QoE of video streaming services becomes paramount to cellular network operators, who need to offer high quality levels to reduce the risks of customers churning for quality dissatisfaction. In this paper, we present a novel approach to multi-dimensional QoE prediction in mobile video using machine learning models. Contrary to previous models for QoE prediction in video streaming, which are generally uni- or low-dimensional and model the impact of single video descriptors independently, we use a high-dimensional input space to model the impact of buffering and initial delay on QoE. We train and test the proposed models on a publicly available mobile video dataset, generated from subjective QoE tests with real viewers. Besides improving prediction performance, the proposed models show that there is a clear influence of other buffering pattern descriptors generally neglected in previous models - in particular those linked to the occurrence of the last stalling event, shedding light on new KPIs to monitor for better QoE assessment in video streaming.**

Keywords*—*QoE; Mobile Video Streaming; Predictive Models; Machine Learning.

I. Introduction

Quality of Experience (QoE) is becoming one of the leading concepts for network management and performance evaluation in operational networks. The intensifying competition among network operators – and in particular in the cellular networks domain, is forcing Internet Service Providers (ISPs) to integrate QoE into the core of their network monitoring and management systems. Among the most relevant QoE-sensitive services consumed by end customers in mobile networks, mobile video takes the pole position. Indeed, mobile video traffic accounts today for more than 60% of the total mobile data traffic [1]. As such, there is an ever-growing interest from cellular network operators to better understand and assess the performance of their networks as perceived by the end users watching video streaming on their devices.

The massive adoption of end-to-end encryption for content distribution - HTTPS, and in particular for video content, motivates the usage of monitoring tools embedded directly at the end points of the service [4]; this allows to directly measure application-layer metrics which are relevant to the performance of the service as perceived by the end user. There are different tools [5], [6], [19] which are capable of monitoring application-layer metrics which are highly correlated to QoE in video streaming services. Buffering events or stallings, video quality/resolution switches and initial playback delay are accepted today as the key application-layer metrics which can be used to predict the QoE undergone by the video watcher, using different models proposed and investigated in the literature [7], [10], [12], [13], [15], [21], [23]–[25]. Out of these metrics, stalling is the paramount one, specially when it comes to mobile video watched in small end-devices such as smartphones; in fact, in [11] we show that QoE for video streaming in modern smartphones is actually slightly impaired by video resolution changes.

While adaptive streaming technology is capable of reducing stalling by reducing video quality and bandwidth requirements, stalling is still the most harmful event for QoE in video streaming. Recent studies show that the occurrence of stalling is far from negligible in operational networks [8], [16], [18], and especially in cellular networks [18], impacting not only the overall user experience, but also user engagement [6], [9]. For these reasons, having a highly accurate model for QoE prediction in mobile video scenarios becomes of capital interest for operators.

In this paper, we explore the possibility of improving previous models for QoE prediction in mobile video by using machine learning. State of the art models are generally uni- or low-dimensional, mapping the number of stallings and the stallings duration to QoE; in general, these models consider only the impact of single video descriptors independently, reducing their accuracy for the sake of simplicity. We present a novel approach to multi-dimensional QoE prediction in mobile video using machine learning models. Contrary to previous models, we use a high-dimensional input space to capture the impact of buffering and initial delay on QoE. We train and test the proposed models on a publicly available mobile video dataset [22], generated from subjective QoE tests with real viewers and multiple stalling patterns. Besides improving prediction performance, the proposed models show that there is a clear influence of other buffering pattern descriptors generally neglected in the past; for example, we show that, in the case of stalling, the location of the final stalling event has a non-negligible impact on QoE, as well as its duration. We benchmark different machine learning based models, compare the best of them to state of the art models, and apply feature

Partly funded by EU project MONROE (H2020-2014-ICT-644399, open call project Mobi-QoE) and by WWTF project Big-DAMA (WWTF-ICT15-129).

analysis and selection techniques to understand the combined impact of different input features describing stalling patterns. Using a decision tree-based regression model, we reduce mean prediction errors by almost 50% as compared to the most accurate state of the art model, i.e., the bi-dimensional, exponential model proposed in [23].

The remainder of the paper is organized as follows: Sec. II presents a brief overview of the related work. Sec. III presents the different ML-based models used for QoE prediction benchmarking, and describes the input features which are extracted directly from the application-layer measurements. Evaluations are reported in Sec. IV, including a characterization of the publicly available subjective QoE measurements used in the modeling exercise, the performance achieved by the proposed models, and a comparison to state-of-the-art models. Features' relevance for QoE prediction is also investigated in this section. Finally, Sec. V concludes this work.

II. Related Work

The problem of QoE assessment in HTTP video streaming is already well-known and well studied, and different QoE models for video streaming have been proposed in the past [7], [10], [12], [13], [15], [21], [23]–[25]. Today it is well accepted that stalling (i.e., stops of the video playback) and initial delay on the video playback are the most relevant KPIs for video streaming QoE [12]–[14], [23]. Quality switches have also a relevant impact on QoE when considering adaptive video streaming technology [15]; however, in [11] we recently found that QoE for video streaming in modern smartphones is actually slightly impaired by resolution switches, mainly due to the screen size of such devices. Recent studies [21] show that the position of stallings and their length have a relevant impact on QoE, but do not attempt to use such metrics to improve QoE predictions. A comprehensive survey of the QoE of adaptive streaming can be found in [3].

Besides pure video quality modeling, other papers [6], [9], [10] have addressed the problem of user engagement prediction for HTTP video streaming.

When it comes to the problem of video streaming QoE measurement in mobile networks, in [17] we introduced the first on-line, large-scale monitoring system for assessing the QoE of YouTube in cellular networks using passive, in-network measurements only. The specific monitoring and assessment of video streaming QoE in smartphones has been also extensively studied in the past [4], [8], [11], [19], [26], [27]; in particular, authors in [26] and [27] tackle the problem of QoE estimation for mobile video streaming apps using machine learning models to map network QoS to QoE.

This paper is complementary to previous work, as it targets the enhancement of previously proposed models for QoE prediction in mobile video streaming, using machine learning to map application-layer metrics to QoE. Different from [26], [27], our study focuses exclusively on the mapping of application-layer metrics to QoE.

III. Machine Learning for Mobile Video QoE

In this paper we propose different Machine Learning (ML) models for QoE prediction, using a high-dimensional feature space as input for the QoE prediction exercise. The proposed model targets the prediction of QoE in HTTP video streaming for mobile scenarios, where smartphones are the default end-device. As we said before and as shown in previous work [11], video quality switches have a limited impact on video streaming QoE in smartphones, given the small screen sizes of such devices. Therefore, the proposed model does not take into account quality switches as input information, and takes as input the stalling pattern and playback delay of a video streaming session. Nevertheless, the modeling approach is generic and can be very easily extended to any other input metrics, including information about quality switches, contextual viewer information, network level QoS measurements, etc.

We assume that a measurement tool such as YoMoApp [19] or similar measures the activity of the video player and reports the occurrence of every single stalling/re-buffering event during a video session, including their duration and exact starting time. We assume that the tool also reports the initial playback delay, which is the time elapsed between the user video request and the actual playback start. Without loss of generality, assuming a video session v_i with n_i stalling events and an initial playback delay d_i, these measurements are reported as two vectors $st_i = \{0, st_i(1), \ldots, st_i(n_i)\}$ and $sd_i = \{d_i, sd_i(1), \ldots, sd_i(n_i)\}$, where st_i reports the starting times of the n_i stalling events and sd_i their corresponding duration. Together, st_i and sd_i fully describe the stalling pattern undergone by video v_i.

In addition, we assume that other video-related metrics such as video duration, frames-per-second, and *content class* (e.g., highly dynamic or mainly static content) are reported in vector vm_i. We note that the metric *content class* is not easy to compute, but include it for the sake of completeness; in any case, we show that the video *content class* has very weak correlation to QoE, and that its impact as part of the model is negligible.

The targeted model should be capable to predict the overall quality as experienced by the end user watching the corresponding video v_i, using as input vector st_i, sd_i and vm_i. In this work we take a standard Mean Opinion Score (MOS) metric to gauge end-user QoE, using an ACR rating scale [2], ranging from "bad" QoE (i.e., MOS = 1) to "excellent" (i.e., MOS = 5) QoE. The corresponding model can be therefore depicted as:

$$\text{MOS}_i = \text{MOS}\left(v_i\right) = \Phi\left(st_i, sd_i, vm_i\right)$$

We use the corresponding input vectors st_i, sd_i, and vm_i to embed video session v_i into a high-dimensional feature space $X\mathbb{R}^m$ where model Φ shall perform the prediction. Tab. I summarizes the set of m input features used for mobile video QoE prediction, which are derived from input vectors st_i, sd_i, and vm_i. The full set includes $m = 19$ different features characterizing the stallings pattern undergone by the video v_i, as well as the particular video contents. Features focus on number and frequency of stalling events, initial playback delay, duration of stallings, as well as their particular location within the video stream.

Model $y = \Phi(X)$ is constructed by learning a specific mapping between input features X and prediction target y,

using a training dataset for which real QoE MOS scores are provided. In this work, we use a publicly available mobile video dataset [22] for model training and validation, generated from subjective QoE tests with real viewers/smartphone users and multiple stalling patterns. Note that model Φ is a regression model, as we assume the most general scenario where the target MOS scores are continuous values in the range $[1, 5]$.

For the sake of benchmarking, we evaluate 11 different machine learning models or *regressors*, all of them well known in the machine learning literature [28]. These include: support vector machines (SVM), decision trees - random tree, bagging-based tree, continuous-prediction tree (M5P) [20], decision stump (DS), discrete regression tree, 3-layers feed-forward neural networks (MLP), random forest (RF), linear and pace regression, and locally-weighted learning (LWL). We use the well-known Weka Machine-Learning software tool[1] to calibrate these models and to perform the evaluations. Parameters are set manually for all the models, performing an extensive trial-and-error testing phase to obtain the best results. We address the interested reader to [28] and to the Weka documentation for additional information on the different configuration parameters of each model.

The finally selected model (M5P) which achieves the best prediction performance is based on decision trees; while decision trees are normally applied in classification problems, it is also possible to use them in regression problems, using different techniques to deal with discretized and missing values [20]. In particular, we adopt the techniques presented in [20], where conventional decision trees are extended with the possibility of performing linear regression at the leaves. Decision trees are a very appealing option; they are simple yet very fast and effective. They are also very easy to interpret, and directly provide filtering rules. In addition, decision trees explicitly show the importance of different features, as the learning algorithm automatically performs feature selection by choosing the most discriminating features. This is a paramount advantage as compared to other ML approaches, as decision trees are more robust to noisy or loosely correlated-to-target input features. Evaluation results are presented next.

IV. Evaluation Results

In this section we compare the performance of the aforementioned regressors using a publicly available subjective QoE measurements dataset. We first describe this dataset and provide a brief overview on the characteristics of the most relevant features. Then we jump into the benchmarking results achieved with the 11 proposed regressors. The best resulting model (M5P) is then compared to 3 different state-of-the-art models proposed in [23]–[25]. Finally, we analyze the relevance of the different input features used in the resulting model for QoE prediction, using group-correlation-based feature selection techniques and intra-features correlation analysis. It is important to note that all ML models are trained and validated by 10-fold cross-validation, reducing overfitting and thus biased results. In a generic k-fold cross-validation approach, the dataset is randomly partitioned into k equal sized sub-sets. Of the k sub-datasets, a single one is used as the validation data for testing the model, and the remaining k-1 sub-sets are used as training data. The cross-validation process is then repeated k times, with each of the k sub-sets used once as validation data. The obtained k results are then averaged to produce a single estimation.

[1] Weka Data Mining, at http://www.cs.waikato.ac.nz/ml/weka/.

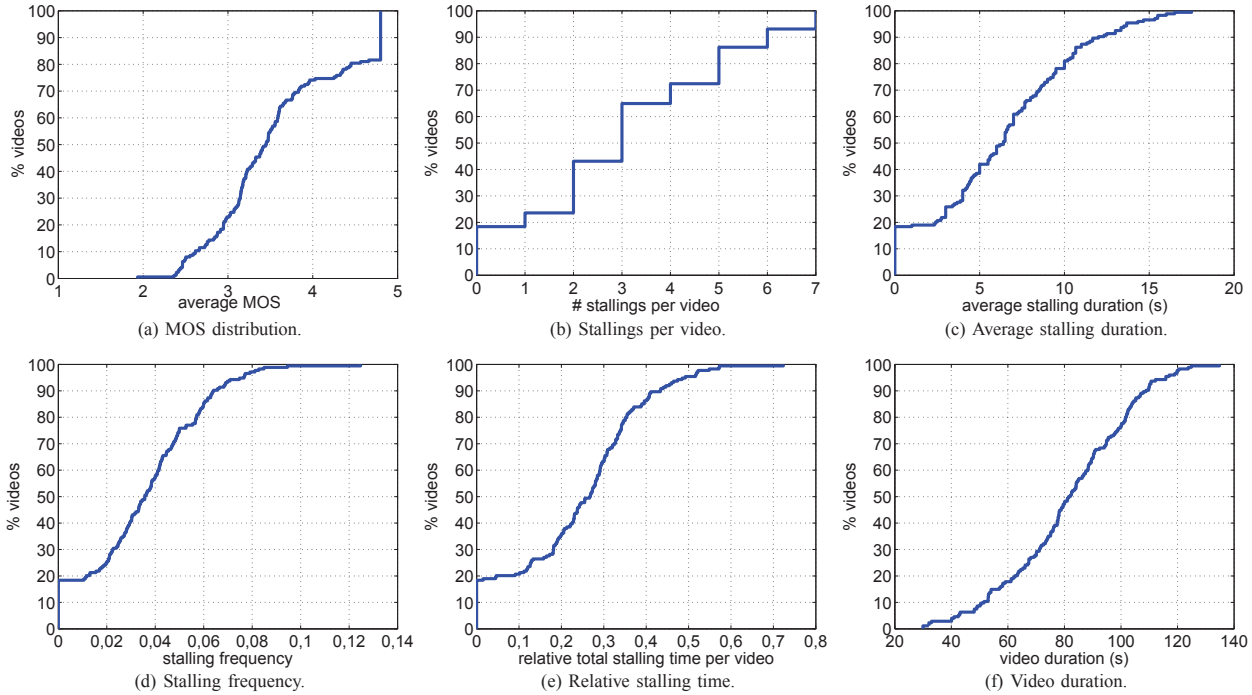
Table I. Input features for mobile video QoE prediction.

	feature	description
f_1	num_stalls	total number of stallings
f_2	freq_stalls	frequency of stallings
f_3	tst	total stalling time
f_4	rel_tst	tst, relative to video duration
f_5	ini_delay	initial playback delay
f_6	rel_ini_delay	ini_delay, relative to video duration
f_7	asd	average stalling duration
f_8	rel_asd	asd, relative to video duration
f_9	t_last_stall_end	elapsed time between end of last stalling and end of the video
f_{10}	sd_last	duration of last stalling
f_{11}	rel_sd_last	sd_last, relative to video duration
f_{12}	loc_last_stall	elapsed time between start of the video and start of last stalling
f_{13}	rel_loc_last_stall	loc_last_stall, relative to video duration
f_{14}	min_sd	minimum stalling duration
f_{15}	median_sd	50%-percentile of stalling duration
f_{16}	max_sd	maximum stalling duration
f_{17}	fps	video frames per second
f_{18}	content_type	video category (e.g., sports, news, etc.)
f_{19}	video_duration	total length of the video
	MOS	average video MOS score

A. Data Description & Characterization

Training and testing of the proposed models is performed on top a publicly available subjective QoE measurements dataset. The LIVE-Avvasi Mobile Video database [22] consists of 174 distorted videos generated from 24 reference videos with 26 unique stalling events and 4830 ratings obtained from 54 subjects who viewed the videos on mobile devices. Reference videos correspond to HD content from YouTube and Vimeo, with a duration range between 29 and 134 seconds (after adding stalling events). Video content spans different categories, including more dynamic contents such as sports to more stable contents such as documentaries, as well as advertisement and music clips. Quality ratings are provided on a standard single stimulus, continuous scale basis, but reported as a Degradation MOS score (DMOS), using a hidden reference removal approach [21]. The dataset includes only the average, per video and per condition DMOS scores. To make results comparable to previous work on video QoE modeling, we re-scale DMOS scores to standard MOS scores, using subjective testing for reference quality assessment. The resulting overall quality is therefore rated from *bad* (i.e., MOS = 1) to *excellent* (i.e., MOS = 5).

Fig. 1 presents a brief characterization of the studied dataset. MOS scores range from 1.9 to 4.8, with about 20% of the videos rated as poor quality (i.e., MOS $<$ 3), 25% rated as good quality (i.e., 3.5 $<$ MOS $<$ 4), and 20% rated as excellent quality (i.e., MOS $>$ 4.5). More than 50% of the tested conditions have 3 or more stalling events, and about 20% of the video conditions consider perfect quality, without



(a) MOS distribution. (b) Stallings per video. (c) Average stalling duration.

(d) Stalling frequency. (e) Relative stalling time. (f) Video duration.

Figure 1. LIVE-Avvasi Mobile Video database description. More than 50% of the tested video conditions have 3 or more stalling events, and about 20% of the video conditions consider perfect quality conditions, without stalling. Average stalling duration ranges mainly from 5 to 15 seconds.

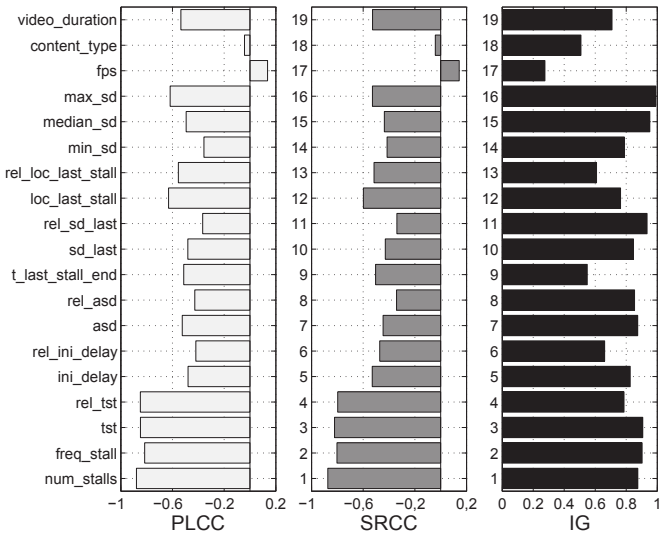


Figure 2. Linear correlation (PLCC), rank correlation (SRCC) and information gain (IG) for input features to MOS scores.

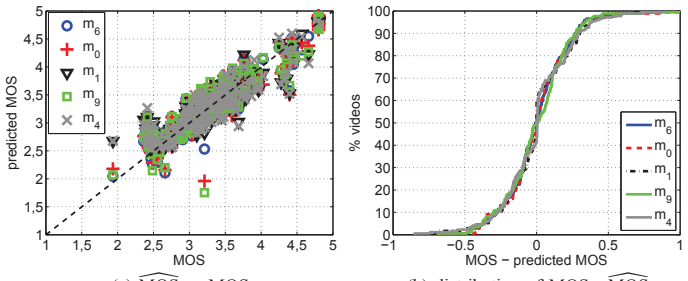
stalling. Stalling events are clearly visible, with ∼55% of the videos having an average stalling duration between 5s and 15s.

To have a first impression on the relevance of each of the 19 input features on QoE, Fig. 2 reports the Pearson linear correlation coefficients (PLCC), the Spearman rank correlation coefficients (SRCC), and the information gain (IG) for each of the 19 input features with respect to the reported QoE MOS scores. As expected, there is a very strong, negative correlation between QoE and number of stallings, frequency of stallings, and total stalling time. The average stalling duration and the initial delay also show strong negative correlation to MOS scores. A very interesting observation is that the location of the last stalling event as well as the duration of the longest stalling event are both highly correlated to MOS scores, suggesting that these metrics could potentially be used to enhance prediction results.

B. Machine Learning Models Performance

We move on now to the evaluation and benchmarking of the proposed models. As we said before, testing and validation are performed following a standard 10-fold cross-validation approach, reducing as such the impact of potential overfitting and biased conclusions. All the 19 input features are considered in these evaluations; feature selection is conducted in Sec. IV-D. Performance is evaluated on the basis of three standard metrics used in regression problems: PLCC coefficients, root mean squared error RMSE = $\sqrt{mean((\hat{X}-X)^2)}$, and mean absolute error MAE = $mean(|\hat{X}-X|)$, where X and $\hat{X}$ are real and predicted values respectively. The MAE metric penalizes all the errors equally, whereas the RMSE metric puts a relatively high weight on larger errors.

Tab. II reports the obtained comparative results for the 11 ML models. Surprisingly, the worst performing model is the MLP, based on neural networks; still, all models achieve a linear correlation above 0.84 and small errors, below 0.31 points in the MOS ACR scale on average. SVM and RF perform particularly well, achieving basically the same results and outperforming almost all the other models. The bagging-based tree model and the pace regression model also achieve an outstanding performance, with similar results. However, the M5P continuous-prediction tree model is the one that



(a) $\widehat{\text{MOS}}$ vs MOS. (b) distribution of MOS - $\widehat{\text{MOS}}$.

Figure 3. Performance of top-5 ML models.

Table II. BENCHMARKING RESULTS FOR ML MODELS.

ID	model	PLCC	RMSE	MAE
m_0	**SVM**	**0.950**	**0.236**	**0.169**
m_1	**RF**	**0.949**	**0.240**	**0.169**
m_2	random tree	0.898	0.344	0.243
m_3	MLP	0.841	0.469	0.255
m_4	**bagging tree**	**0.946**	**0.246**	**0.176**
m_5	LWL	0.852	0.397	0.297
m_6	**M5P**	**0.957**	**0.220**	**0.167**
m_7	linear regression	0.878	0.362	0.310
m_8	additive regression DS	0.921	0.301	0.231
m_9	**pace regression**	**0.948**	**0.242**	**0.179**
m_{10}	discrete regression tree	0.911	0.315	0.224

achieves the best results, with a linear correlation of almost 0.96, average absolute error below 0.17 and a RMSE = 0.22.

Fig. 3 provides more details on the results achieved by the top-5 models, depicting in (a) the predicted MOS scores vs. the real ones and (b) the distribution of the absolute errors, MOS - $\widehat{\text{MOS}}$. The 5 models perform similarly, without incurring in higher errors in particular regions of the MOS scale. Indeed, errors look equally distributed and equally sized along the complete scale; in addition, as observed in the distribution of errors in (b), all the models tend to very slightly underestimate the MOS scores, and more than 80% of the video sessions QoE are estimated with an absolute error below 0.25. A negligible fraction of the sessions' QoE values are estimated with absolute errors above 0.5. Based on the aforementioned nice properties offered by decision trees, we select the M5P decision tree model as the underlying model for QoE prediction.

Fig. 4 depicts an approximated version of the corresponding M5P model, using binning on the prediction target MOS scores. Note that in this case the tree leaves seem to overlap, which shall be interpreted as only for visualization purposes. The binning process actually provides non-overlapping leaves. The first node of the tree corresponds to the location of the last stalling event, relative to the video duration, showing its relevance. Indeed, when stallings happen at the very beginning of the video session - i.e., in the first 15% of the total video playback time, QoE is mainly defined by the initial playback delay and the total stalling time, relative to the video duration. This is coherent with the observations done in [24], [25], in which memory and recency effects have a key role in the overall quality perception of the viewer. On the contrary, when the last stalling event occurs in the complement, the QoE is mainly defined by the number of stallings and their total, relative duration.

Table III. BENCHMARKING RESULTS FOR M5P AND STATE OF THE ART MODELS.

model	PLCC	RMSE	MAE
M5P	0.957	0.220	0.167
exp. (original) [23]	0.844	1.516	1.361
exp. (fit)	0.907	0.437	0.323
HW [24]	0.891	–	–
DQS [25]	0.864	0.300	–

The finally obtained M5P model can be transformed into a linearized stump-like tree with 2 leaves, in which each leave represents a linear mapping between the most relevant inputs and MOS scores, and the split is done on the basis of number of stallings (f_1). The model can be expressed as:

$$\text{MOS} = \begin{cases} \sum_{i=1}^{6} \alpha_i \times F(i) + \alpha_0, & \text{if } f_1 \leq 1.5 \\ \sum_{i=1}^{6} \beta_i \times F(i) + \beta_0, & \text{if } f_1 > 1.5 \end{cases}$$

where $F = \{f_4, f_1, f_8, f_6, f_9, f_{17}\}$ correspond to the most relevant features selected by the M5P model, which include relative total stalling time, number of stallings, relative initial delay and average stalling duration, elapsed time since the end of the last stalling event till end of the video, and video frames per second, the latter with an almost negligible weight α_6, β_6.

C. Benchmarking to SotA Prediction Models

We compare now the performance achieved by the selected M5P model, with that achieved by 3 different state-of-the-art models proposed in [23]–[25]. The model presented in [23] is one of the most cited models in the video streaming QoE literature, and corresponds to an exponential-based, bi-dimensional model, which maps the number of stallings and their average duration into a MOS score. We compare two different variants of this model: *exp. (original)* corresponds to the exponential model using as parameters the default values presented in [23]; to be fair to this model, we additionally test a re-calibrated version referred to as *exp. (fit)*, where model parameters are fit to the used dataset. The model presented in [24] is a more complex model, which accounts for non-linearities and recency effects, taking as input not only the number and duration of stallings, but also their position in the video stream. The model uses a standard Hammerstein-Wiener non-linear filter model with memory to capture recency and hysteresis effects of human perception, and is calibrated in the same dataset used in this paper. We refer to this model as *HW* model. Finally, the model presented in [25] uses a state-machine-based model to account for the cumulative impact of stallings and increased satisfaction during normal playback. We refer to this model as *DQS* (Delivery Quality Score).

Tab. III reports the obtained results. The table only reports partial results for the HW and DQS models, which come

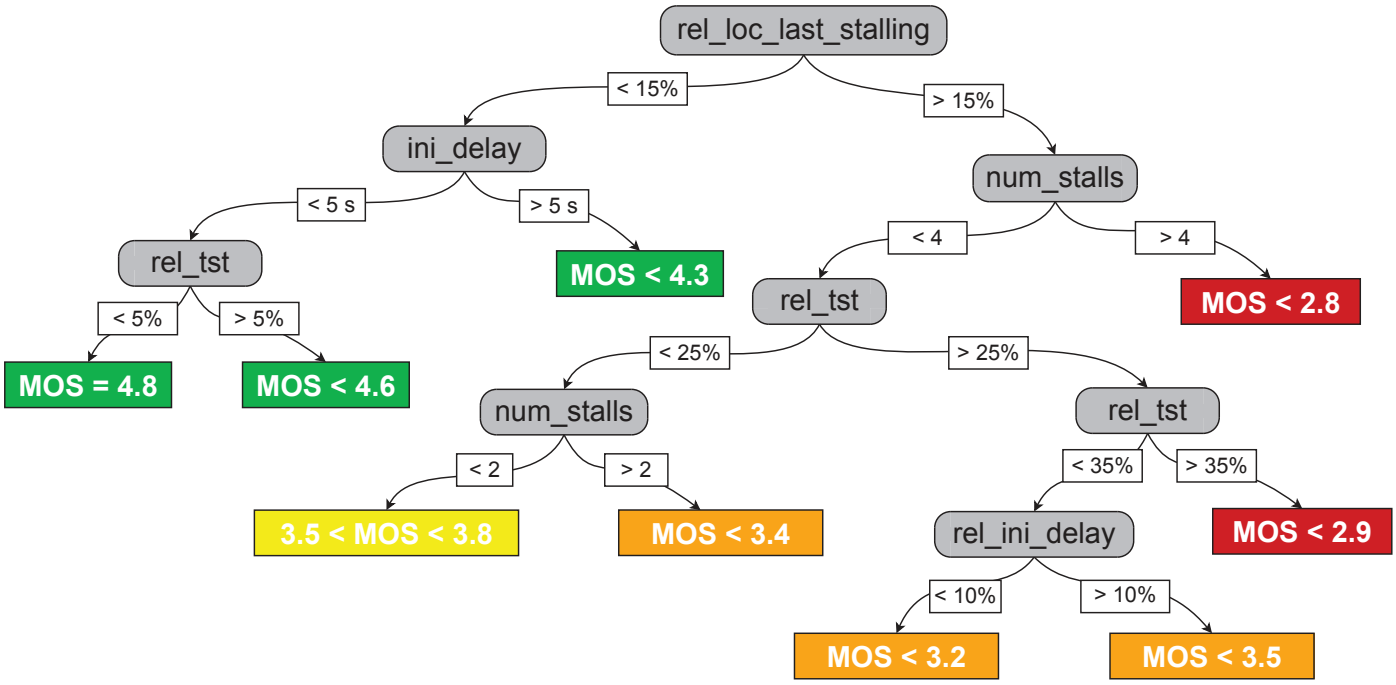
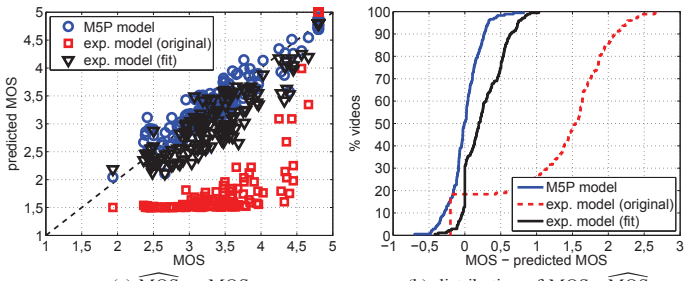


Figure 4. Decision tree for MOS prediction - approximation based on discretization, without using linear regression at the leaves.



(a) $\widehat{\text{MOS}}$ vs MOS. (b) distribution of MOS - $\widehat{\text{MOS}}$.

Figure 5. Performance of M5P and exponential, state of the art models.

directly from the obtained results in [24], [25]. The M5P model clearly outperforms state-of-the-art models, providing a much higher linear correlation to the actual MOS scores. The exponential model comes in the second place, after a careful re-calibration of its underlying parameters. Still, prediction errors for the *exp. (fit)* model are as large as twice the errors achieved by the proposed M5P model. Fig. 5 depicts (a) the predicted MOS scores vs. the real ones and (b) the distribution of the absolute errors incurred by the M5P model and the *exp.* model, using both original and fit parameters. Results achieved with the original exponential model are very poor, as this model systematically underestimates QoE; the re-calibrated version of the model achieves much better results, but still suffers from underestimation bias. We can therefore conclude that the proposed M5P model predicts QoE better than state-of-the-art models.

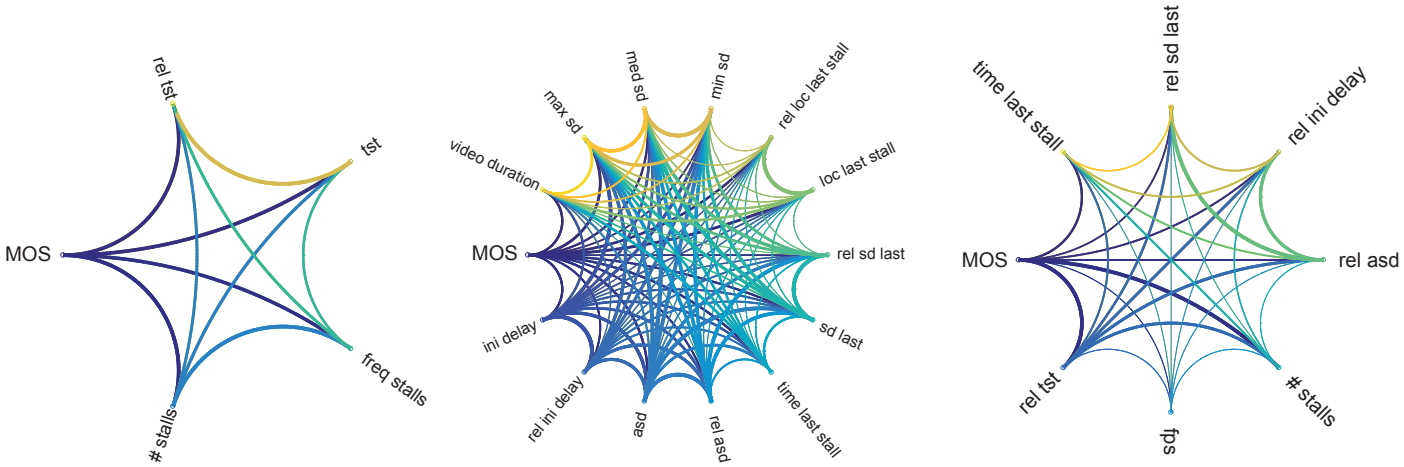
D. Feature Space Analysis

To conclude the study, we present a very brief analysis on the impact of the different input features on QoE. Fig. 6 plots the inter-features PLCC coefficients and linear correlation to MOS scores considering (a) the top-4 strongest correlated to MOS features, (b) the remaining 13 features describing a stalling-pattern, and (c) the top-6 features automatically selected by group-correlated-based feature selection. Colors are only used to distinguish features, but the thicker the edge connecting two nodes, the higher the PLCC coefficient value.

Besides reflecting both a strong correlation to MOS scores and among input features, it is very interesting to verify that automatic feature selection is capable to select those features which basically describe the M5P model. In particular, we see that, even if some of the selected features have an almost negligible one-to-one correlation to MOS scores - e.g., such as video frames per second (fps), they play a role within the model achieving the best prediction performance, i.e., the M5P one.

V. Concluding Remarks

In this paper we have introduced a novel machine learning based model for multi-dimensional QoE prediction in mobile video streaming. Based on decision trees, the proposed model outperforms previously proposed state-of-the-art models by reducing prediction errors between 25% and almost 50%. Besides improving prediction performance, the proposed model shows that there is a clear influence of other stalling pattern descriptors generally neglected in previous models - in particular those linked to memory effects and the occurrence of the last stalling event. As such, this model permits to enhance current measurement tools and systems for video streaming QoE prediction, suggesting novel metrics to measure in the future.



(a) strong PLCC correlated features (b) mean PLCC correlated features (c) group-correlated-based selected features

Figure 6. Circular plots reflecting inter-features PLCC coefficients and linear correlation to MOS scores. In (c), group-correlated-based feature selection results in 7 features which include the F features describing the M5P model.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper," 2017. [Online].

[2] Int. Telecommunication Union, "ITU-T Rec. P.800: Methods for Subjective Determination of Transmission Quality," 1996.

[3] M. Seufert et al., "A Survey on Quality of Experience of HTTP Adaptive Streaming", in *IEEE Comm. Surveys & Tutorials*, vol. 17, no. 1, 2015.

[4] P. Casas et al., "Next to You: Monitoring Quality of Experience in Cellular Networks from the End-devices", in *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, 2016.

[5] B. Staehle et al., "YoMo: A YouTube Application Comfort Monitoring Tool," in *QoEMCS*, 2010.

[6] H. Nam et al., "YouSlow: What Influences User Abandonment Behavior for Internet Video?," Tech. report 2017. [Online].

[7] A. Balachandran et al., "Developing a Predictive Model of QoE for Internet Video", in *ACM SIGCOMM*, 2013.

[8] Q. Chen et al., "QoE Doctor: Diagnosing Mobile App QoE with Automated UI Control and Cross-layer Analysis", in *ACM IMC*, 2014.

[9] F. Dobrian et al., "Understanding the Impact of Video Quality on User Engagement", in *ACM SIGCOMM*, 2011.

[10] S. Krishnan et al., "Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-experimental Designs", in *ACM IMC*, 2012.

[11] P. Casas et al., "Exploring QoE in Cellular Networks: How Much Bandwidth do you Need for Popular Smartphone Apps?", in *ACM All Things Cellular Workshop*, 2015.

[12] T. Hoßfeld et al., "Quantification of YouTube QoE via Crowdsourcing", in *IEEE International Symposium on Multimedia*, 2011.

[13] R.K.P. Mok et al., "Inferring the QoE of HTTP Video Streaming from User-Viewing Activities", in *ACM W-MUST Workshop*, 2011.

[14] T. Hoßfeld et al., "Initial Delay vs. Interruptions: Between the Devil and the Deep Blue Sea", in *QoMEX*, 2012.

[15] T. Hoßfeld et al., "Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming", in *QoMEX*, 2014.

[16] J. Jiang et al., "Shedding Light on the Structure of Internet Video Quality Problems in the Wild", in *ACM CoNEXT*, 2013.

[17] P. Casas et al., "YOUQMON: A System for On-line Monitoring of YouTube QoE in Operational 3G Networks", *ACM SIGMETRICS PER*, vol. 41, 2013.

[18] P. Casas et al., "An Educated Guess on QoE in Operational Networks through Large-Scale Measurements", in *ACM Internet-QoE Workshop*, 2016.

[19] F. Wamser et al., "Understanding YouTube QoE in Cellular Networks with YoMoApp – a QoE Monitoring Tool for YouTube Mobile", in *ACM MOBICOM*, 2015.

[20] Y. Wang et al., "Induction of Model Trees for Predicting Continuous Classes", in *ECML*, 1997.

[21] D. Ghadiyaram et al., "Study of the Effects of Stalling Events on the Quality of Experience of Mobile Streaming Videos", in*IEEE GlobalSIP*, 2014.

[22] D. Ghadiyaram et al., "LIVE Mobile Stall Video Database,", [online]: http://live.ece.utexas.edu/research/LIVEStallStudy/index.html, 2016.

[23] T. Hoßfield et al, "Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience", in *Data Traffic Monitoring and Analysis: From measurement, classification and anomaly detection to Quality of Experience* 7754, 264–301, Springer Berlin Heidelberg, 2013.

[24] D. Ghadiyaram et al., "A Time-varying Subjective Quality Model for Mobile Streaming Videos with Stalling Events", in *SPIE Optical Engineering Applications*, 2015.

[25] H. Yeganeh et al., "Delivery Quality Score Model for Internet Video", in *IEEE International Conference on Image Processing*, 2014.

[26] P. Casas et al., "Predicting QoE in Cellular Networks using Machine Learning and in-Smartphone Measurements", in *QoMEX*, 2017.

[27] V. Aggarwal et al., "Prometheus: Toward Quality-of-Experience Estimation for Mobile Apps from Passive Network Measurements", in *ACM HotMobile*, 2014.

[28] R.O. Duda et al., "Pattern Classification, 2nd Edition", Wiley, 2000.