QoE beyond the MOS: Added Value Using Quantiles and Distributions

Tobias Hoßfeld*, Poul E. Heegaard†, Martin Varela‡
*University of Duisburg-Essen, Modeling of Adaptive Systems, Essen, Germany
Email: tobias.hossfeld@uni-due.de
†Department of Telematics, Norwegian University of Science and Technology (NTNU)
Email: poulh@item.ntnu.no
‡ VTT Technical Research Centre of Finland
Email: Martin.Varela@vtt.fi

***Abstract*—Traditionally, Quality of Experience (QoE) assessment results (or objective estimations thereof) are presented as a single scalar value, typically a Mean Opinion Score (MOS). While useful, the limitations of MOS are evident even in its name; for many applications, just having a mean value is simply not enough. For service providers in particular, it would be more interesting to have an idea of how the scores are distributed, so as to ensure that a certain portion of the user population is experiencing satisfactory levels of quality, thus reducing churn. In this paper we propose different statistical measures to express important aspects of QoE beyond MOS like user diversity, uncertainty of user rating distributions, ratio of dissatisfied users. Further, we propose a way to use MOS values and the standard deviation of opinion scores (SOS) hypothesis, which postulates a quadratic relation between subjective scores and their standard deviation, in order to derive quantiles for subjective ratings.**

I. Introduction

It is a common and well-established practice to use Mean Opinion Scores (MOS) [1] to quantify perceived quality, both in the research literature, as well as in practical applications such as QoE models. This is simple and useful for "technical" evaluation of systems and applications such as network dimensiong, performance evaluation of new networking mechansims, assessment of new codecs, etc.

From a service provider's point of view, however, MOS values are not sufficient. Averages only consider — well — averages, and do not provide a way to address variations between users. As an extreme example, if the MOS of a given service under a given condition is 3, it is impossible to know whether all users perceived quality as acceptable (all scores are 3), or maybe half the users rated the quality 5 while the other half rated it 1, or anything in between, in principle. To some extent, this can be mitigated by quantifying user rating variation via e.g. standard deviations. However, the question often faced by service providers is of the type: "Assuming they observe comparable conditions, are at least 95% of my users satisfied with the service quality they receive?". Our goal in this paper is to provide a way to answer such questions.

Our main contribution is highlighting the importance of insight in the uncertainty of the opinion scores. The uncertainty is hidden by the MOS, and such an insight will enable the service providers to manage the QoE in a much better way. The paper shows different approaches to quantify the uncertainty; standard deviation, entropy, cumulative density function (CDF), and quantiles. An example with a fitting of data to a binomial distribution, and another example where the uncertainty is estimated directly from the data, are presented.

The remainder of this paper is structured as follows. Section II discusses related work. Section III describes our approach, while Section IV-A presents the results of applying it to two subjective assessment dataset, hightlighting the potential and limitations of the approach. We conclude the paper in Section V, discussing the practical implications of our results.

II. Background and Related Work

A. User Diversity and Assessment Methodologies

In the context of traditional multimedia quality assessment, there exist several methodologies that have been developed and standardized [2]–[4] in order to obtain quality ratings for different types of media and applications. These methodologies have been designed with consistency and reproducibility in mind, which allow results to be comparable across studies done in similar conditions. For the most part, these methodologies result in MOS ratings, along with standard deviation and confidence intervals, which is in line with recommendations in the statistics literature, e.g. [5].

The use of averaging, however, can hide information about the way different users rate the quality that is useful in many contexts. Such limitations of MOS are recently discussed [6], [7]; assessment results should contain more details on the variation of opinions among users [8] . There are many potential reasons for these variations, first and foremost the fact that users are simply different people, with different expectations, service usage history, etc. Even though these factors can be, to some degree, controlled in a lab environment, in real applications, these different opinions can have a significant meaning in terms of business (e.g., service provider reputation and branding image, customer acquisition, churn).

In [6], the authors proposed a compact way of capturing some of the user variation in the assessments by means of combining the MOS with the standard deviation of opinion

Acknowledgment: This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant HO 4770/1-2 (DFG Oekonet).

scores, which provides a first step towards improving the traditional approaches.

B. Service Provider's Interest in Quantiles

In order to stay in business in a free market, ISPs and other service providers need to maintain a large portion of their users satisfied, lest they stop using the service or change providers — the dreaded "churn" problem. For any given service level the provider can furnish, there will be a certain proportion of users who might find it unacceptable, and the perceived quality of the service is one of the key factors determining user churn [9]. Moreover, a large majority ($\sim 90\%$) of users will simply defect a service provider without even complaining to them about service quality, and report their bad experience within their social circles [10], resulting in a possibly even larger business impact in terms of e.g., brand reputation. With only a mean value as an indicator for QoE, such as the MOS, the service provider cannot know what this number of unsatisfied users might be, as user variation is lost in the averaging process.

For many applications, however, it is desirable to gauge the portion of users that is satisfied given a set of conditions (e.g., under peak-time traffic, for an IPTV service). For example, a service provider might want to ensure that at least, say, 95% of its users find the service acceptable or better. In order to ascertain this, some knowledge of how the user ratings are distributed for any given condition is needed. In particular, calculating the 95% quantile (keeping in line with the example above) would be sufficient for the provider.

III. Theory: Distribution of Opinion Scores

For the sake of simplicity, but without loss of generality, we consider a discrete rating scale with values from $0, 1, \dots, N$. In the QoE domain, the most commonly used scale for quality ratings is a discrete 5-point scale with the categories $1 \triangleq$'bad', $2 \triangleq$'poor', $3 \triangleq$'fair', $4 \triangleq$'good', and $5 \triangleq$'excellent' referred to as Absolute Category Rating (ACR) [11]. In the numerical results in Section IV-A and Section IV-B, we analyze subjective results based on this 5-point ACR scale and shift the ratings accordingly to $0, 1, 2, 3, 4$ with $N = 4$.

In what follows, we briefly introduce the notation used in the paper, and summarized in Table I. We consider a subjective test with r participants for a particular test condition. Thus, we obtain r ratings on the ACR scale. U is the random variable that represents the opinion score, $U \in \{0, \dots, 4\}$. The probability mass function, $p_u = P(U = u)$, is the probability that the opinion score is u. An unbiased estimate of p_u is the ratio of users that are rating the test condition with u. With the rating of user i being U_i, we obtain $\hat{p}_u = \frac{1}{r}\sum_{i=1}^{r} \delta_{U_i,u}$ with the Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

The MOS score is $x = \sum_{u=0}^{N} u \cdot \hat{p}_u$, and the expected score value is $\mathrm{E}[U] = \sum_{u=0}^{N} u \cdot p_u$. A *dissatisfied user* is defined as a user that has a score value below a certain threshold value θ. The probability of a dissatisfied user is then $\mathbb{P}_\theta = P(U \leq \theta) = \sum_{i=0}^{\lfloor \theta \rfloor} p_u$. The probability of a dissatisfied user is estimated by $\hat{\mathbb{P}}_\theta = \sum_{u=0}^{\lfloor \theta \rfloor} \hat{p}_u$.

TABLE I: Variables and notations frequently used in the paper.

notation	meaning
N	upper number on the discrete rating scale $\{0, \dots, N\}$
U	random variable describing the user ratings for a particular test condition, $U \in \{0, \dots, N\}$
r	number of user ratings per test condition
U_i	rating of user i for a particular test condition
p_u	probability that the user rating is $u \in \{0, \dots, N\}$
$\hat{p}_u$	estimate of p_u, i.e. ratio of users who rate the test condition with u
x	MOS value for a test condition
θ	acceptance threshold, $0 \leq \theta \leq N$, used in the definition of *dissatisfied users*, i.e. $U_i < \theta$
$\mathbb{P}_\theta$	probability of an dissatisfied user with threshold θ,
$\hat{\mathbb{P}}_\theta$	estimate of the probability $\mathbb{P}_\theta$ of an dissatisfied user,
$V(x)$	variance of user ratings as a function of MOS scores x
a	SOS parameter of SOS hypothesis in Eq. 2
ΔQ_α	difference between the MOS value and the α-quantile
$\Delta\Theta_\theta$	difference between the MOS value and the threshold θ

A. SOS Hypothesis Revisited

On a discrete rating scale, there is a maximum standard deviation score $S^*(x)$ related to a certain MOS value x which is the result of diverse ratings of the same test condition from r different users. The following equation holds for any MOS value $x \in [0; N]$, cf. [6].

$$S^*(x) = \sqrt{-x^2 + N \cdot x} \tag{1}$$

The SOS hypothesis formulates a generic relationship between MOS and SOS values independent of the type of service or application under consideration. Thereby, the SOS parameter a is a specific value for a certain application or service (and the test conditions/impairments under consideration) which is derived from subjective tests. For a given MOS value x on a discrete rating scale from 0 to N, the standard deviation is

$$S(x) = \sqrt{a(-x^2 + N \cdot x)}\,. \tag{2}$$

B. Entropy

The entropy is defined as [12]

$$E = -\sum_{u=0}^{N} p_u \log p_u\,. \tag{3}$$

The entropy expresses the uncertainty in the information provided by the measurement system. In a perfect score system, where all factors of the users and the state of the system are known, the entropy is 0 because the score will be deterministic. When nothing is known, and all score values have the same probability, the entropy is at its maximum. The uncertainty of the measurement system should in principle be the same over the whole range of values, x, but due to the discrete and truncated scale that is used, the effect of the uncertainty is less evident close to the maximum and minimum values.

C. User Ratings following a Binomial Distribution

Now we consider the cases that user ratings U follow a binomial distribution for a particular test condition. As reported in [6], (and in the case study in Section IV-A), the measurements $\tilde{U}$ can be described by a binomial distributed random variable $U \sim BINO(N,p)$. Thereby, the user ratings can take values $0, 1, \cdots, N$, while the parameter p is computed based on the observed MOS value x as measured for that test condition.

The expected value of $U \sim BINO(N,p)$ is

$$\mathrm{E}\left[U\right] = N \cdot p = x\,. \tag{4}$$

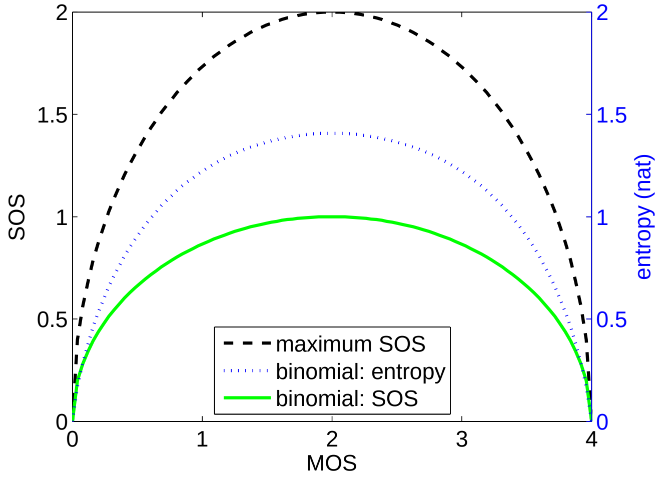


Fig. 1: Relationship between MOS and SOS values. The maximum possible SOS as well as the SOS for binomially distributed user ratings are depicted. The entropy for the binomial distribution depending on the MOS is given on the right y-axis.

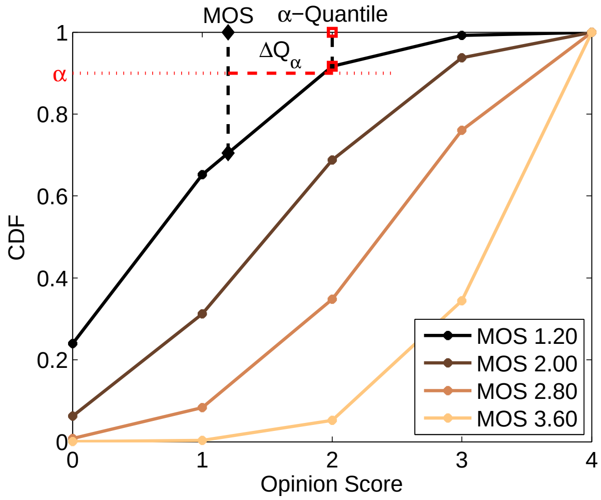


Fig. 2: The cumulative distribution function (CDF) for different test conditions is given for user ratings following a binomial distribution.

From Eq.(4), the parameter p can be computed $p = {}^{x}\!/\!{}_{N}$ for any given x.

The standard deviation of $U \sim BINO(N,p)$ is

$$\mathrm{STD}\left[U\right] = \sqrt{N \cdot p \cdot (1-p)} = \sqrt{x - \frac{x^2}{N}}\,. \tag{5}$$

We now take a closer look at the relation between SOS and of MOS in Eq.(5) and compare it with the SOS hypothesis in Eq.(2).

$$\sqrt{a(-x^2+Nx)} = \sqrt{x - \frac{x^2}{N}} \tag{6}$$

Thus, for binomially distributed user ratings, then $a = \frac{1}{N}$, which implies $a = 0.25$ when $N = 4$.

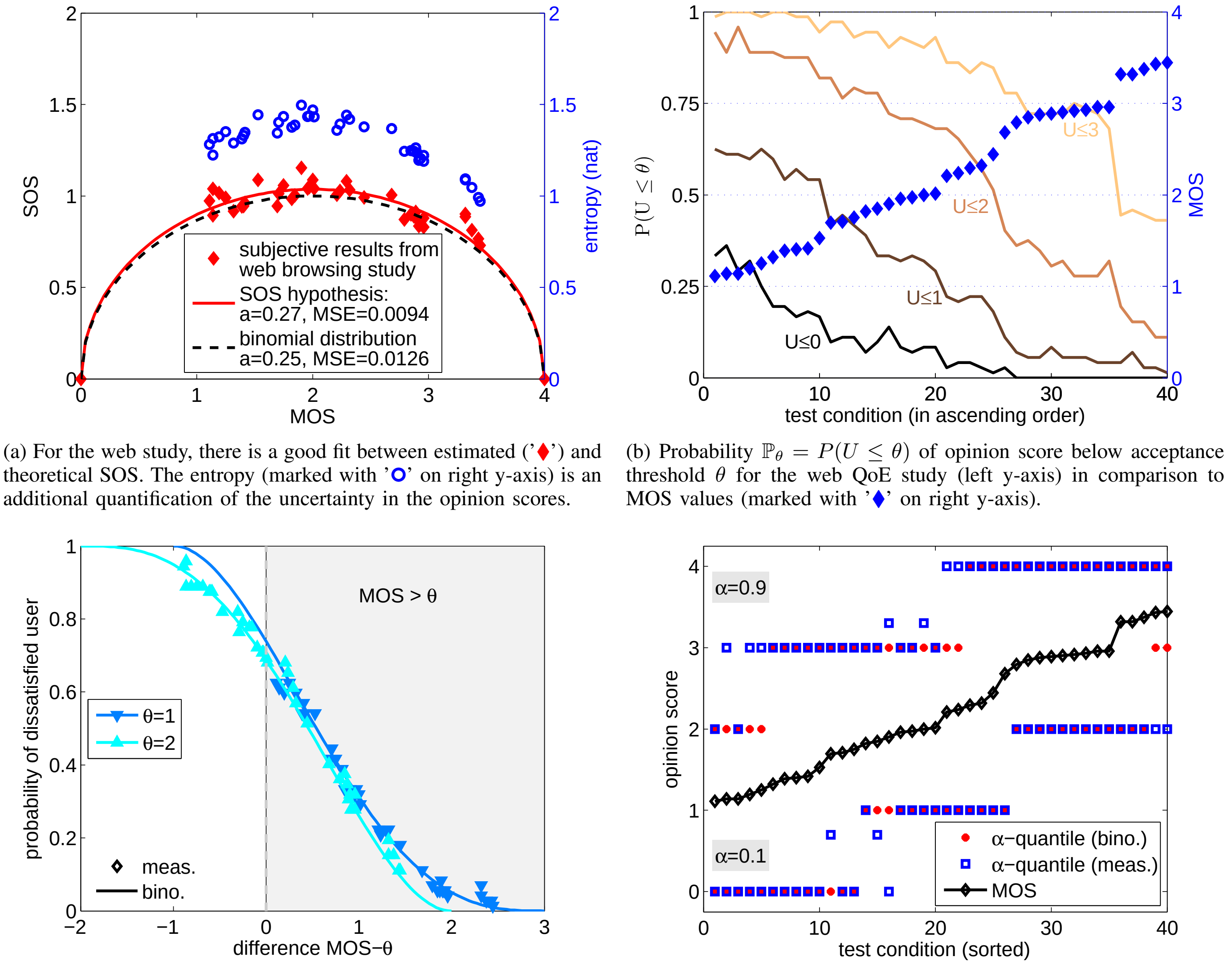
Figure 1 shows the relationship between MOS and SOS. For any given MOS x, there exists a maximum possible SOS as a result of the discrete rating scale. Further, the relationship between MOS and SOS is shown for user ratings following a binomial distribution $a = 0.25$ on a 5-point rating scale. Each test condition will lead to one point on the curve, i.e. a tuple of MOS and SOS value. The subjective results of the web QoE study in Section IV-A lead to a SOS parameter of $\tilde{a} = 0.27$. Here, the binomial distribution can be used to describe the entire distribution of user ratings for any obtained MOS value x. The curve for entropy has the same shape as the SOS curve, as most uncertainty about the measurement system and user ratings is given in the middle of the rating scale.

For binomially distributed user ratings, Figure 2 shows the cumulative distribution function (CDF) of the user ratings U for different MOS values $x \in \{1.2, 2.0, 2.8, 3.6\}$. Each test condition will lead to one instantiation of the binomial distribution with parameters corresponding to the MOS value observed for that test condition. Beside the MOS, the α-quantile is given for $\alpha = 0.9$. While the MOS value represents the average user rating, the quantile quantifies the opinion score that the fraction α of users observes at most. In the next sections, we will investigate the difference between MOS and quantiles for subjective studies on web and video QoE, and discuss how to apply the quantiles to add value to the management of QoE.

IV. Application to Real Data Sets

A. Example Use Case: Web QoE

We consider in the following the results from a subjective user study on web QoE which is based on the experiments in [13]. Subjects were browsing a set of web pages while the page loading times (PLT) were delayed to quantify the impact of PLT on web QoE. The participant interacted with a Java applet that already contained the contents of the websites. The applet simulated the download of various web pages with predefined page load times. The web page also contained rating buttons from 1 to 5, which were used by the test user to give his/her personal opinion score during the browsing session. In particular, subjects were asked to answer the question “Are you satisfied with this download speed?”.



(a) For the web study, there is a good fit between estimated ('♦') and theoretical SOS. The entropy (marked with '○' on right y-axis) is an additional quantification of the uncertainty in the opinion scores.

(b) Probability $\mathbb{P}_\theta = P(U \le \theta)$ of opinion score below acceptance threshold θ for the web QoE study (left y-axis) in comparison to MOS values (marked with '♦' on right y-axis).

(c) With MOS above the threshold, $MOS > \theta$ for the web browsing study, the ratio of dissatisfied users, $\mathbb{P}_\theta$ is significant.

(d) The $10\,\%$- and $90\,\%$- quantiles for the web browsing study are significantly different from the MOS.

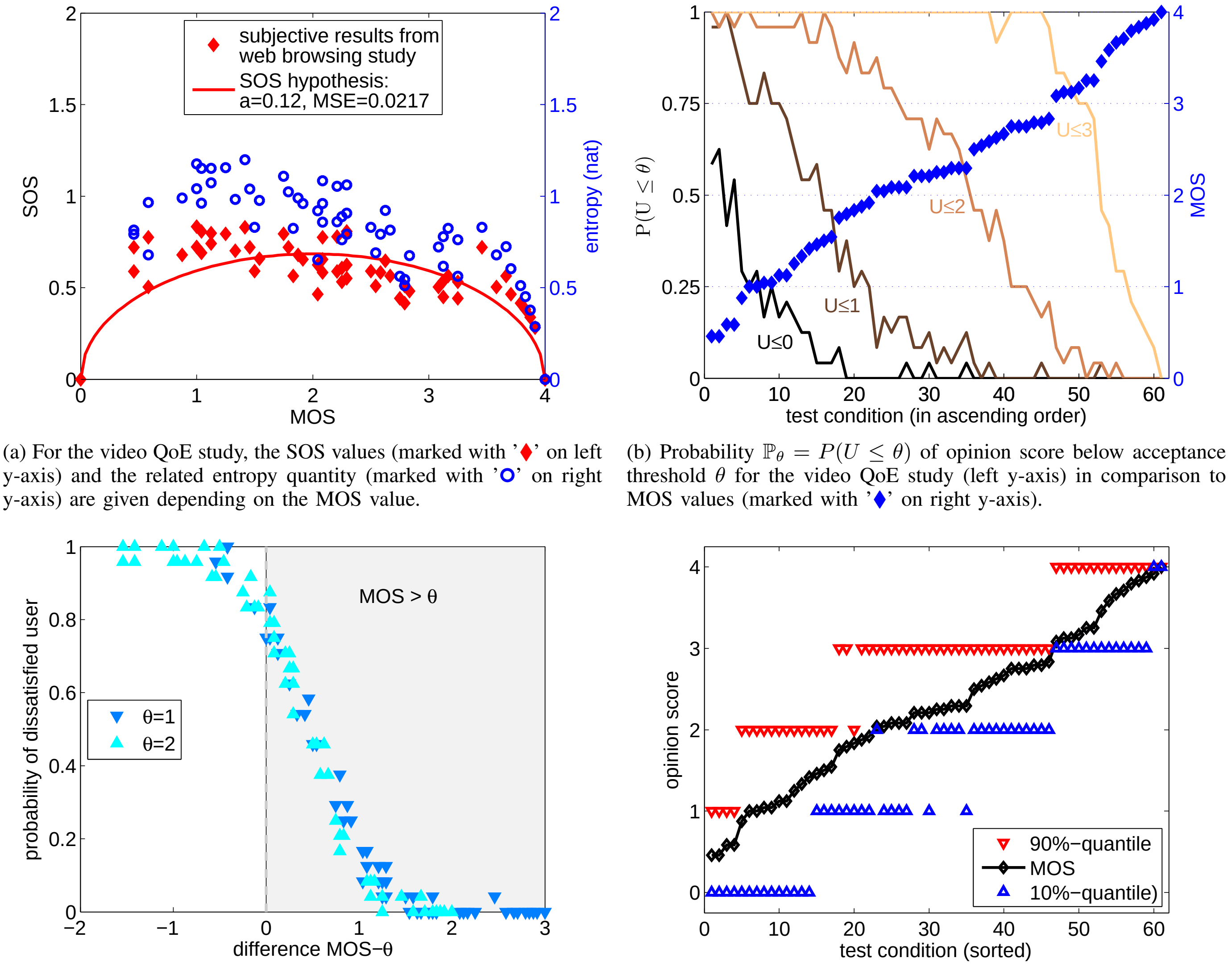
Fig. 3: Web browsing study [13]: The page load time was influenced for each test condition and 72 subjects rated the overall QoE. Each user viewed various web pages with different PLTs resulting into 40 test conditions per user.

During the tests, a user viewed 40 web pages. There were 72 users completing the online test. The maximum PLT was only 1.2 s in order not to scare the users away due to long waiting times and an accordingly frustrating user experience. The minimum and the mean PLT were 0.24 s and 0.66 s.

Figure 3a shows the relationship between SOS and MOS and reveals the diversity in user ratings. The tuple of MOS and SOS from the subjective ratings is plotted with a red diamond for each test condition. In addition, the entropy for each test condition is given, which quantifies the uncertainty in the measurement system and the unpredictability of the user ratings. This means that even for a given MOS the individual user rating is relatively unpredictable due to the user rating diversity which is also expressed by the higher entropy values.

The results in Figure 3a confirm the SOS hypothesis and the SOS parameter is obtained by minimizing the least squared error between the subjective data and Eq. 2. As a result, a SOS parameter of $\tilde{a} = 0.27$ is obtained. The mean squarred error (MSE) between the subjective data and the SOS hypothesis (solid curve) is close to zero, indicating a very good match. In addition, the MOS-SOS relationship for the binomial distribution $(a = 0.25)$ is plotted as dashed line. It can be seen that the results can be approximated nicely by the binomial distribution (MSE=0.01). Thus, the theoretical results in Section III-C are valid for this web QoE study too.

Figure 3b uncovers the averaging effect of the MOS which is not sufficient to fully understand the results for the subjective study. In particular, the empirical probability of an



(a) For the video QoE study, the SOS values (marked with '♦' on left y-axis) and the related entropy quantity (marked with '○' on right y-axis) are given depending on the MOS value.

(b) Probability $\mathbb{P}_\theta = P(U \leq \theta)$ of opinion score below acceptance threshold θ for the video QoE study (left y-axis) in comparison to MOS values (marked with '♦' on right y-axis).

(c) With MOS above the threshold, $MOS > \theta$ for the video QoE study,, the ratio of dissatisfied users, $\mathbb{P}_\theta$ is significant.

(d) The 10 %- and 90 %- quantiles for the video QoE study are significantly different from the MOS.

Fig. 4: Video QoE study [14]: Video quality was affected by resolution, amount of motion and loss process in the network. For this study, 24 subjects rated audiovisual, then audio and video quality for IPTV-like streams. These results correspond to the video quality ratings for 61 conditions.

dissatisfied user is related to the observed MOS values (plotted as diamonds). It can be seen that even for good overall MOS values, a significant number of users perceives the quality as bad or very bad.

In Figure 3c the probability of a dissatisfied user, $\mathbb{P}_\theta$, is plotted for two threshold values, $\theta = 1$ and $\theta = 2$ for the binomial distribution. The corresponding MOS values are also included for each of the 40 web tests, and the threshold values are indicated. From the plot it is apparent that even the MOS is above the threshold, a significant amount of the users are dissatisfied, measured as the probability of dissatisfied user, $\mathbb{P}_\theta$. E.g., with threshold values $\theta = 1$ and $\theta = 2$, for all tests where the MOS are above the threshold, the $\mathbb{P}_\theta$ is as much as 60% when the MOS are close to the threshold, and even as much as 30% when $\mathrm{MOS} - \theta \approx 1$.

This information, while very significant to service providers, is masked out by the MOS. Thus, the entire distribution, the ratio of dissatisfied users or other measures need to be reported. While the SOS values reflect the user diversity, and the entropy the measurement uncertainty, the quantiles help to understand the fraction of users with very bad (e.g. 10 % quantile) or very good quality perception (e.g. 90 % quantile).

Figure 3d shows the measured quantiles (as well as the quantiles from the binomial distribution) compared to the MOS values for $\alpha = 0.1$ and $\alpha = 0.9$, respectively. The difference between the MOS value and the α-quantile is significant, with a maximum difference of 1.5 on the rating scale. Thus, a fraction α of users rates an opinion score which is up to 1.5 away from the MOS. The MOS hides this information which may be very valuable and important for the service provider. Quantiles of subjective user ratings will give additional relevant information and we recommend to provide the 10 %- and the 90 %-quantiles to fully understand the meaning of the results. By means of the SOS parameter a, appropriate distributions can be selected for the fitting, like the binomial distribution in case of the web QoE results with $a = 0.25$. Then, the entire information (i.e. CDF, quantiles, ratio of dissatisfied users, entropy) can be derived.

B. Example Use Case: Video QoE

In this section, we consider a portion of the results presented in [14], where an audiovisual model for IPTV-like streaming was proposed. In the study, 24 users assessed first audiovisual, then audio and video quality under a variety of conditions (among which 61 effectively different ones for video). The parameters considered included the video resolution, amount of movement, and the loss process (loss rate and burstiness) in the network.

In contrast to the results for web QoE, the user ratings do not follow a binomial distribution but show lower variances, see Figure 4a. We further observe that the SOS values are more spread around the SOS hypothesis curve. This is an indicator that important influential factors are not included in the measurements. In this experiment, this might be because the video and audio were assessed at the same time. This can also be seen from the entropy values. The uncertainty in the measurements system depends less on the MOS values, which implies that not all factors (user and system related) are captured. Therefore, we recommend to provide information about SOS-MOS relationships and entropy.

Similarily to the web experiments, we take a closer look at the probability of an dissatisfied user for the video experiments in Figure 4b and compare it with the MOS value. Again, it can be also observed that the MOS values hide relevant information for the service provider about the ratio of dissatisfied users. Figure 4c visualizes the probability that a user is dissatisfied (i.e. $\mathbb{P}_\theta = P(U \leq \theta)$, although the MOS value exceeds this threshold. The results show that up to 60 % of users are happy, although the MOS is larger than θ. To quantify the ratio of users at the edge, it is therefore recommended to provide quantiles as in Figure 4d. For the video experiments, no simple distribution can be used to fit the user ratings. In such a case, it is recommended to estimate the quantiles and CDF from the dataset to get a better view on the user ratings.

V. Conclusions

In this paper we state the case for using different quantities for describing the results of QoE assessments. These quantities give more insights and are of especial importance for service providers to understand QoE. We recommend to use:

1, MOS: average user rating for one test condition
2, SOS: user diversity for that test condition
3, Quantile: user rating of fraction of (satisfied, dissatisfied) users close to the acceptance threshold, θ
4, Entropy: the uncertainty of the measurement system and the unpredictability of individual user ratings
5, Probability distribution: complete information about the randomness in the opinion score.

For many experiments, a compact description of the results is possible via the SOS parameter a. This allows to rebuild the entire distribution as we have demonstrated for the web QoE experiments. There, the distribution of user ratings could be described by the binomial distribution with $a = 0.25$. With the complete information from the distribution, other measures beyond MOS like quantiles can be computed to get significant insights relevant for service providers. This has been demonstrated for dissatisfied users where a significant large number of users was not satisfied, although the overall MOS score was above a certain threshold. For cases where the assessments don't follow a binomial distribution, the quantiles or CDF can be estimated from the dataset.

The availability of these results is important to better understand some business-related aspects of services, such as the ratio of dissatisfied users, which is useful for predicting churn rates. Likewise, going beyond the MOS allows service providers to better provision their services so that a target fraction of the user population is satisfied.

References

[1] ITU-T, “ Recommendation P.800.1 – Mean Opinion Score (MOS) terminology,” July 2006.

[2] International Telecommunication Union, “Mean Opinion Score (MOS) Terminology,” *ITU-T Recommendation P.800.1*, Mar. 2003.

[3] International Telecommunication Union, “Subjective video quality assessment methods for multimedia applications,” *ITU-T Recommendation P.910*, April 2008.

[4] International Telecommunication Union, “Methodology for the subjective assessment of the quality of television pictures,” *ITU-R Recommendation BT.500-12*, Mar. 2009.

[5] Warren S. Torgerson, *Theory and Methods of Scaling*, John Wiley & Sons,Inc., 1963.

[6] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger, “SOS: The MOS is not enough!,” in *QoMEX 2011*, Mechelen, Belgium, Sept. 2011.

[7] Robert C Streijl, Stefan Winkler, and David S Hands, “Mean opinion score (mos) revisited: methods and applications, limitations and alternatives,” *Multimedia Systems*, pp. 1–15, 2014.

[8] Evangelos Karapanos, Jean-Bernard Martens, and Marc Hassenzahl, “Accounting for diversity in subjective judgments,” in *Proc of 27th International Conference on Human factors in Computing Systems*, New York, NY, USA, 2009, CHI ’09, pp. 639–648, ACM.

[9] Hee-Su Kim and Choong-Han Yoon, “Determinants of subscriber churn and customer loyalty in the korean mobile telephony market,” *Telecommunications Policy*, vol. 28, no. 9âĂŞ10, pp. 751 – 765, 2004.

[10] D. Soldani, M. Li, and R. Cuny, *QoS and QoE Management in UMTS Cellular Systems*, Wiley, 2006.

[11] Sebastian Moeller, *Assessment and Prediction of Speech Quality in Telecommunications*, Springer, 1st edition, August 2000.

[12] C. E. Shannon, “A mathematical theory of communication,” *Bell system technical journal*, vol. 27, 1948.

[13] Tobias Hoßfeld, Raimund Schatz, Sebastian Biedermann, Alexander Platzer, Sebastian Egger, and Markus Fiedler, “The Memory Effect and Its Implications on Web QoE Modeling,” in *23rd International Teletraffic Congress (ITC 23)*, San Francisco, USA, Sept. 2011.

[14] T. Mäki, D. Kukolj, D. Đorđević, and M. Varela, “A Reduced-Reference Parametric Model for Audiovisual Quality of IPTV Services,” in *QoMEX*, July 2013.