Formal Definition of QoE Metrics

Tobias Hoßfeld*, Poul E. Heegaard†, Martín Varela‡, Sebastian Möller§

*Modeling of Adaptive Systems, University of Duisburg-Essen, Germany, Email: tobias.hossfeld@uni-due.de
† Norwegian University of Science and Technology (NTNU), Email: poul.heegaard@item.ntnu.no
‡ VTT Technical Research Centre of Finland — Communication Systems, Email: Martin.Varela@vtt.fi
§ Quality and Usability Lab, TU Berlin, Germany, Email: Sebastian.Moeller@telekom.de

Abstract—This technical report formally defines the QoE metrics which are introduced and discussed in the article "QoE Beyond the MOS: An In-Depth Look at QoE via Better Metrics and their Relation to MOS" by Tobias Hoßfeld, Poul E. Heegaard, Martín Varela, Sebastian Möller, accepted for publication in the Springer journal 'Quality and User Experience' [2]. Matlab scripts for computing the QoE metrics for given data sets are available in GitHub [3].

I. Definition of QoE Metrics

A. Preamble

We consider studies where users are asked their opinion on the overall quality (QoE) of a specific service. The *subjects* (the participants in a study that represent users), rate the quality as a *quality rating* on a *quality rating scale*. As a result, we obtain an *opinion score* by interpreting the results on the rating scale numerically. An example 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 an Absolute Category Rating (ACR) scale [6].

Let U be a random variable (RV) that represents the quality ratings, $U \in \Omega$, where Ω is the rating scale, which is also the state space of the random variable U. The RV U can be either discrete, with probability mass function f_u, or continuous, with probability density function $f(u)$.

The notation used in the paper is summarized in Table I.

B. Observations in a study are samples of U

In a study, each observation can be regarded as samples $U_{i,r,c}$ of U. Each sample corresponds to the observation i of subject r, under test condition c, using the rating scale Ω. Typically, the notation can be simplified, e.g., if the identity of a subject is not an issue, and we look at only one test with the same test conditions c, then the set of observations will form a set of samples $\mathcal{U} = \{U_i\}$, $i = 1, \cdots, R$. More details about test conditions are found in the next section.

C. Specification of test conditions

A study typically specifies a set of test conditions, $\mathcal{C}$, which includes both technical and non-technical factors. The set of test conditions contains an invariant (static) part that is unchanged over all tests, $\mathcal{C}_0$, and a variable (dynamic) part that changes between each of the J tests, $\mathcal{C}_j$, $(j = 1, \cdots, J)$. For each test, the test conditions might change according to a pattern specified on a set (of size k), $\mathcal{C}_j = \{c_{j,1}, \cdots, c_{j,k}\}$. If $\mathcal{C}_j$ is not the same for all R subjects in test j, then the notation

must be extended and indexed with, r, i.e. $\mathcal{C}_{j,r}$. Furthermore, in some studies we don't have the same number of subjects per test condition. To specify this, you should add an index to the number of subjects R_j of test j.

As an example, consider a study of the effect of a specific sequence of changes in the video quality classes, where each subject is expected to rate the overall technical quality of the video. In each test a specific set of quality classes, q_n, is defined, let's say N classes, in an unordered set $\mathcal{Q} = \{q_1, \cdots, q_N\}$. The variable part of $\mathcal{C}$ is then defined as an ordered set $\mathcal{C}_j = (c_{j,1}, \cdots, c_{j,k})$, where c contains both the quality class, q, and the time, t, for the quality class changes, $c = (t, q)$. If the sequence of changes is not the same for each subject in a test, either with respect to q or t, then the notation must capture this by extending and indexing the $\mathcal{C}_j$ with r as described above. If the test is about the effect of a number of changes, k, and not a specific sequence of such, then $\mathcal{C}_j = (k, \mathcal{Q})$. Finally, if each test only has one condition,

Table I
KEY VARIABLES AND NOTATIONS USED IN THE PAPER.

notation	meaning
MOS	Mean Opinion Score
SOS	Standard deviation of Opinion Score
ACR	Absolute Category Rating
U	random variable (RV) for quality ratings
U^+	upper limit of a rating scale
U^-	lower limit of a rating scale
$\mathcal{U}$	set of ratings $\{U_i\}$ for a particular test condition
R	number of subjects (= ratings) per test condition
J	number of tests
$\mathbb{S}$	Statistical definition set ($\{\Omega, \mathcal{C}, \Sigma, \mathcal{S}\}$)
Ω	quality rating scale (same as the sample space of the RV U)
$\mathcal{C}$	set of test conditions
Σ	set of statistics
$\mathcal{S}$	set of (sufficient) observators
u	quality rating value for a test condition (e.g. MOS on the ACR scale)
f_u	probability that the quality rating is $u \in \Omega$
$\hat{f}_u$	estimate of f_u, i.e. ratio of subjects who rate the test condition with u
S(u)	SOS as a function of mean opinion score u
a	SOS parameter of SOS hypothesis in Eq. (11)
$\mathbb{A}_\theta$	α-acceptability, i.e. the probability that ratings are above θ, $P(U > \theta)$
%PoW	Poor-or-Worse (in %)
%GoB	Good-or-Better (in %)
%TME	Terminate Early (in %)

Table II

Table II
DIFFERENT EXAMPLES OF TEST CONDITION SETS FOR THE CASE WITH THE CHANGES IN THE VIDEO QUALITY CLASSES.

Test conditions	Description of test condition for test j
$C_j = \{q_j\}$	Fixed quality level ($k = 1$)
$C_j = \{k, \mathcal{Q}\}$	k random changes of quality level
$C_j = \{(t_1, q_1), \cdots, (t_k, q_k)\}$	k specific changes in quality level
$C_{j,r}$	specific changes for each subject r

then $\mathcal{C}_j = c_j$, where $k = 1$ and $c_j \in \mathcal{Q}$.

See Table II for a summary of the different options in this example.

To give an illustration of how to use this in a real study, let's consider the example of web QoE [2]. Here we have $R = 72$ subjects who were downloading $k = 40$ web pages, with different content w_i, and page loading time (PLT), q_i, which gives $c_j = (w_j, q_j)$. The test conditions can be specified either as a sequence of $k = 40$ content and impairments, with $C_1 = \{(w_1, q_1), \cdots, (w_{40}, q_{40})\}$, or alternatively, if the specific sequence doesn't matter, then the test can be described by $J = 40$ separate tests with one test condition each, $C_j = \{(w_j, q_j)\}$, $j = 1, \cdots, 40$. In this study, all subjects, r, had the same $\mathcal{C}$.

D. n-tuple statistical definition set

An experimental design of a study needs to specify both technical and non-technical factors, and the kind of statistics that will be obtained. To make sure that all necessary details about the statistics are included, it can be useful to have a compact notation for this. In this paper we propose an n-tuple, $\mathbb{S} = \{\Omega, \mathcal{C}, \Sigma, \mathcal{S}\}$ for this purpose, where

- Ω is the *rating scale*, i.e. the sample set, $U_i \in \Omega$, e.g. $\{1, \cdots, 5\}$, e.g. $[0; 6]$.
- $\mathcal{C}$ is the set of test conditions[1].
- Σ is the statistics that will be obtained, e.g. MOS, SOS, median, quantiles, etc.
- $\mathcal{S}$ is the sufficient observers for the estimators of the statistics in Σ, e.g. $\sum U_i$, $\sum U_i^2$, etc.

The notation here should be considered as an example of a minimal n-tuple. Necessary extensions might apply.

E. Expected value and its estimate: MOS

The expected value (sometime referred to as the *mean*) of the random variable, is

$$E[U] = \begin{cases} \sum_{u=U^-}^{U^+} u f_u & : U \text{ is discrete} \\ \int_{u=U^-}^{U^+} u f(u) du & : U \text{ is continuous} \end{cases} \quad (1)$$

This is an example of a 1st-order statistics, along with the *median* and the *mode*.

The Mean Opinion Score (MOS) is an estimate $\hat{U}$ of $E[U]$.

$$\text{u} = \text{MOS} = \hat{U} = \sum_{v=1}^{N} v \hat{f}_v \quad (2)$$

[1]the static part $\mathcal{C}_0$ can be defined outside the n-tuple

where $\hat{f}_u$ is the estimated probability of opinion score u,

$$\hat{f}_u = \frac{1}{R} \sum_{i=1}^{R} \delta_{U_i, u} \quad (3)$$

with the Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. Substituting $\hat{f}_u$ from (3) in (2) gives,

$$\hat{U} = \sum_{v=1}^{N} v \hat{f}_v = \sum_{v=1}^{N} v \frac{1}{R} \sum_{i=1}^{R} \delta_{U_i, v} \quad (4)$$

Now, let $U_i = \sum_{v=0}^{N} v \delta_{U_i, v}$, and by reorder the sums, then

$$\hat{U} = \frac{1}{R} \sum_{i=1}^{R} \sum_{v=1}^{N} v \delta_{U_i, v} = \frac{1}{R} \sum_{i=1}^{R} U_i \quad (5)$$

Hence, we have two, unbiased estimators of $E[U]$,

$$\hat{U} = \sum_{v=1}^{N} v \hat{f}_v = \frac{1}{R} \sum_{i=1}^{R} U_i \quad (6)$$

F. Standard deviation and its estimate: SOS

The variance σ_U^2 of U is

$$Var[U] = \begin{cases} \sum_{u=U^-}^{U^+} (u - E[U])^2 f_u & : U \text{ is discrete} \\ \int_{u=U^-}^{U^+} (u - E[U])^2 f(u) du & : U \text{ is continuous} \end{cases} \quad (7)$$

The standard deviation of U is $\sigma_U = \sqrt{Var[U]}$. This is an example of a 2nd-order statistics, which contains more information than the mean, since it gives a measure of the uncertainty of the mean.

The standard deviation of the opinion score (SOS), is an estimate of the σ_U:

$$S = \sqrt{\frac{1}{R-1} \sum_{i=1}^{R} U_i^2 - \frac{R}{R-1} \hat{U}^2} \quad (8)$$

The SOS is estimated over the sample set $\mathcal{U} = \{U_i\}$ taken from either a discrete or continuous distribution.

It is important to note that the S is the *estimated standard deviation of the opinion score*, U, which converges to $S \to \sigma_U$ for a large number of samples, R. This must not be confused with the *standard error* of the $\hat{U}$, which is s.e.$(\hat{U}) = S/\sqrt{R}$ and converges to $S/\sqrt{R} \to 0$ for a large R.

G. SOS as function of MOS

In [4], the minimum, $S^-(u)$, and the maximum SOS, $S^+(u)$ were obtained, as a function of the MOS=u. The minimum SOS is $S^-(u) = 0$ on a continuous scale, $[U^-; U^+]$, and

$$S^-(u) = \sqrt{u(2\lfloor u \rfloor + 1) - \lfloor u \rfloor(\lfloor u \rfloor + 1) - u^2} \quad (9)$$

on a discrete scale, $\{U^- \cdots U^+\}$.

The maximum SOS is, on both continuous and discrete scales (the scales as above),

$$S^+(u) = \sqrt{-u^2 + (U^- + U^+)u - U^- \cdot U^+} \quad (10)$$

The SOS hypothesis [4], formulates a generic relationship between MOS and SOS values independent of the type of service or application under consideration.

$$S(u) = \sqrt{a} \cdot S^+(u) \qquad (11)$$

It has to be noted that the SOS parameter a is scale invariant when linearly transforming the user ratings and computing MOS and SOS values for the transformed ratings. The SOS parameter allows to compare user ratings across various rating scales. Thus, any linear transformation of the user ratings does not affect the SOS parameter a which is formally proven in Section II-A. However, it has to be clearly noted that if the participants are exposed to different scales, then different SOS parameters may be observed. The parameter $a \in [0;1]$, depends on the application or service, and the test conditions, and is derived from subjective tests. The computation of the SOS parameter is provided Eq.(35) in Section II-B.

H. Quantiles and Distribution of User Ratings

The state space Ω of the random variable U can be either discrete, with probability mass function f_u, or continuous, with probability density function $f(u)$. If an observed set $\mathcal{U} = \{U_i\}$ from a study contains a large number of observations, then the f_u or $f(u)$ can be estimated (the most well know probability density estimation approach is the histogram). For most studies in the QoE domain, estimating *quantiles* is an alternative option that can be quite useful.

The n'th q-quantile of a random variable U is a value $Q_{n/q}$, such that the probability $P(U \leq Q_n/q) \leq n/q$ and $P(U \geq Q_n/q) \leq (q-n)/q$, $(0 < n < q)$. Well-know quantiles are

- $q = 2$: this is called the *median*
- $q = 4$: this is called the *quartile*
- $q = 100$: this is called the *percentile*

Assume that we have an ordered set of R quality ratings $\mathcal{U} = \{U_i\}$ from a study. The q-quantiles for the ratings can then be estimated by

$$\hat{Q}_{n/q} = \mathcal{U}^{(h)} \qquad (12)$$

where $h = \lceil Rn/q \rceil$, $(0 < n < q)^2$.

As shown above, this is also called the α'th percentiles, with $\alpha = 10$ and 90, for our examples.

I. θ-Acceptability

For service providers, acceptance is an important metric to plan, dimension and operate their services. Therefore, we would like to establish a link between opinion measurements from subjective QoE studies and behavioral measurements. In particular, it would be very useful to derive the "accept" behavioral measure from opinion measurements of existing QoE studies. This would allow to reinterpret existing QoE studies from a business oriented perspective. Therefore, we introduce the notion of θ-acceptability which is based on opinion scores.

The θ-acceptability, $\mathbb{A}_\theta$, is defined as the probability that the opinion score is above a certain threshold θ, $P(U \geq \theta)$,

---

²other estimators for $\hat{Q}_q$ exists

---

and can be estimated by $\hat{f}_s$ from Eq. (3) or by counting all user ratings $U_i \geq \theta$ out of the R ratings.

$$\mathbb{A}_\theta = \int_{s=\theta}^{U^+} \hat{f}_s ds = \frac{1}{R} |\{U_i \geq \theta : i = 1, \ldots, R\}| \qquad (13)$$

J. Acceptance

When a subject is asked to rate the quality as either *acceptable* or *not acceptable*, this means that U is *Bernoulli*-distributed. The quality ratings are then samples of $U_i \in \{0, 1\}$, where $1 \triangleq$ 'accepted' and $0 \triangleq$ 'not accepted'. The probability of acceptance is then $f_u = P(U = u)$, $U \in \{0, 1\}$, and can be estimated by (3) with $u = 1$:

$$\hat{f}_1 = \frac{1}{R} \sum_{i=1}^{R} \delta_{U_i, 1} \qquad (14)$$

K. %GoB, %PoW, and %TME

In the past, service providers have also based their planning on (estimated) percentages of users judging a service as "poor or worse" (%PoW), "good or better" (%GoB), or the percentage of users abandoning a service (Terminate Early, %TME). These percentages have been calculated from MOS distributions on the basis of large collections of subjective test data, or of customer surveys. Whereas the original source data is proprietary in most cases, the resulting distributions and transformation laws have been published in some instances. One of the first service providers to do this was Bellcore [5], who provided transformation laws between an intermediate variable, called the Transmission Rating R, and %PoW, %GoB and %TME. These transformation were further extended to other customer behavior predictions, like retrial (to use the service again) and complaints (to the service provider). The Transmission Rating could further be linked to MOS predictions, and in this way a link between MOS, %PoW and %GoB could be established. The E-model, a parametric model for planning speech telephony networks, took up this idea and slightly modified the Transmission Rating calculation and the transformation rules between R and MOS, see [1]. The resulting links can be seen in Fig. 1. Such links can be used for estimating the percentage of dissatisfied users from the ratings of a subjective laboratory test; there is, however, no guarantee that similar numbers would be observed with the real service in the field. In addition, the subjective data the links are based on mostly stem from the 1970-1980s; establishing such links anew, and for new types of services, is thus highly desirable.

The measures *Poor-or-Worse* (%PoW), *Good-or-Better* (%GoB), and *Terminate Early* (%TME), are all quantile levels in the distribution of the quality rating U, or in the empirical distribution of $\mathcal{U} = \{U_i\}$.

In the E-model [1] the RV of the quality rating, $U \in [0; 100]$ (referred to as the Transmission Rating R), represents an objective (estimated) rating of the voice quality. It is assumed that $U \sim N(0, 1)$, which is the standard normal distribution, with the cumulative distribution function,

$$F_U(u) = P(U \leq u) = \int_{-\infty}^{u} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt \qquad (15)$$

Table III
MOS AND TRANSMISSION RATING R WITH THE QUANTILE MEASURES.

MOS	R	%PoW	%GoB	%TME
1.00000	6.52	99.192	0.041	96.732
1.50000	27.27	86.611	2.039	70.736
1.87293	36.00	71.311	6.681	50.000
2.00000	38.68	65.349	9.139	43.340
2.31513	45.00	50.000	17.425	28.689
2.50000	48.57	41.176	23.747	21.608
3.00000	58.08	20.685	45.221	8.381
3.10000	60.00	17.425	50.000	6.681
3.50000	67.96	7.563	69.062	2.288
4.00000	79.37	1.585	88.699	0.336
4.50000	100.00	0.029	99.379	0.003
5.00000	undefined	0.000	100.000	0.000

Under this assumption, the measures have been defined as[3]

$$\text{GoB}(u) = F_U(\frac{u-60}{16}) = P_U(U \geq 60) \quad (16)$$

$$\text{PoW}(u) = F_U(\frac{45-u}{16}) = P_U(U \leq 45) \quad (17)$$

$$\text{TME}(u) = F_U(\frac{36-u}{16}) = P_U(U \leq 36) \quad (18)$$

(the % of the measures are obtained by multiplying each of the measures by 100).

The E-model also defines a transformation of the U to a continuous scale MOS$\in [1; 4.5]$, by:

$$MOS(u) = 7 \cdot (u-60) \cdot (100-u) \cdot u \cdot 10^{-6} + 0.035 \cdot u + 1 \quad (19)$$

By use of this transformation, the three measures are plotted for (continuous) MOS ($\in [1; 4.5]$) in Figures 1. Observe that the sum of %GoB + %PoW do not add up to 100%, because the probability (denoted "neutral" in the figure), $P(40 < U < 60)$, is not included in either %PoW or %GoB. The quantiles used (i.e. 36, 45 and 60) for the three measures, and the assumed standard normal distribution, are chosen as a result of a large number of subjective tests conducted while developing the E-model [1]. Table III includes the MOS and the Transmission Rating R, with their corresponding values of the %PoW, %GoB, and %TME measures. Observe: all values of MOS on the ACR scale are included, even for MOS=5 where the Transmission Rating R is not defined. Furthermore, the table contains the MOS values (real numbers) for the quantiles for each of the three measures, i.e., when $R = 36, 45, 60$.

The measures are estimated based on the ordered set of quality ratings, $\mathcal{U} = \{U^{(i)}\}$, by using the θ-Acceptability estimator from Eq. (13). First, discretise the quality rating scale $\mathcal{U} \in \{0, 100\}$. Then, using the Eq. (13), the following applies

$$\%\hat{\text{GoB}} = \mathbb{A}_{\theta_{gb}} \quad (20)$$

$$\%\hat{\text{PoW}} = 1 - \mathbb{A}_{\theta_{pw}} \quad (21)$$

$$\%\hat{\text{TME}} = 1 - \mathbb{A}_{\theta_{te}} \quad (22)$$

[3]When $U \sim N(0,1)$ then $F_U(u) = 1 - F_U(-u)$, which is applied for the GoB definition.

For example, in the E-model the $\theta_{gb} = 60$, $\theta_{pw} = 45$, and $\theta_{te} = 36$ for $\mathcal{U} \in \{0, 100\}$, and $\theta_{gb} = 3.1$, $\theta_{pw} = 2.3$, and $\theta_{pw} = 1.9$ on a $\mathcal{U} \in \{1, 5\}$ scale (when using Eq. 15).

$$\%\hat{\text{GoB}} = \mathbb{A}_{\lceil 3.1 \rceil} = \mathbb{A}_4 = P(U \geq 4) \quad (23)$$

$$\%\hat{\text{PoW}} = 1 - \mathbb{A}_{\lfloor 2.31513 \rfloor} = 1 - \mathbb{A}_2 = P(U \leq 2) \quad (24)$$

$$\%\hat{\text{TME}} = 1 - \mathbb{A}_{\lfloor 1.87293 \rfloor} = 1 - \mathbb{A}_1 = P(U \leq 1) \quad (25)$$

It is important to note that the quantiles in the examples are valid for *speech quality tests* under the assumptions given in the E-model. The mapping of MOS to the %PoW and %GoB metrics in Table III are specific for this E-model, but the %PoW and %GoB metrics are general and can be obtained from any quality study, provided that the thresholds θ_{gb} and θ_{pw} are determined.

The purpose of the example above is to demonstrate GoB, PoW, TME using an ACR scale (1-5). This is a theoretical exercise (valid for the E-model) where we apply the transformation from R to "MOS" (term used when E-model was introduced) as given in Eq. (19), and transform Eq. (16)-(18) into Eq. (20)-(22), using the notation introduced in Table III. Samples from Eq. (19) are given in Table III. The %GoB = $P(R \geq 60)$ corresponds to %GoB = $P(\text{MOS} \geq 3.1)$ which on an integer scale is %GoB = $P(\text{MOS} \geq 4)$. Correspondingly, for %PoW = $P(R \leq 45) = P(\text{MOS} \leq 2.32) = P(\text{MOS} \leq 2)$.
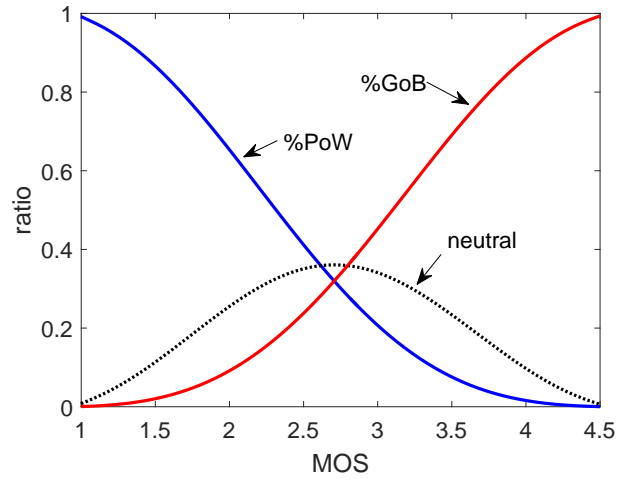

Figure 1. Relationship between MOS, %PoW and %GoB as used in the E-model [1]. The ratio of users not rating poor or worse as well as good or better is referred to as 'neutral' and is computed by $1 - \%\text{GoB} - \%\text{PoW}$.

II. DETAILS ABOUT COMPUTATION OF SOS PARAMETER

A. Invariance of SOS Parameter for Linearly Transformed User Ratings

In a subjective experiment, we observe the random variable (RV) U for quality ratings (for a certain, fixed test condition). In the experiment, a continuous rating scale is used with lower bound L_1 and higher bound H_1, i.e. $U \in [L_1; H_1]$. We observe the SOS parameter a.

The expected value of U is $E[U] = u$ with variance $Var[U] = V_1(u)$ according to the SOS hypothesis with

$$V_1(u) = a_1(-u^2 + (L_1 + H_1)u - L_1 \cdot H_1). \quad (26)$$

Now, the user ratings U are linearly transformed to another rating scale $[L_2; H_2]$ by the transformation function

$$\tau(U) = \frac{U - L_1}{H_1 - L_1}(H_2 - L_2) + L_2. \quad (27)$$

The variance of the transformed user ratings is

$$
\begin{aligned}
Var[\tau(U)] &= Var\left[\frac{U - L_1}{H_1 - L_1}(H_2 - L_2) + L_2\right] && (28) \\
&= Var\left[\frac{H_2 - L_2}{H_1 - L_1}U\right] && (29) \\
&= \left(\frac{H_2 - L_2}{H_1 - L_1}\right)^2 \cdot Var[U]. && (30)
\end{aligned}
$$

However, the latter term is equivalent to the variance according to the SOS hypothesis with SOS parameter a on the transformed rating scale, i.e.

$$V_2(\tau(u)) = a_2(-\tau(u)^2 + (L_2 + H_2)\tau(u) - L_2 \cdot H_2). \quad (31)$$

For the user ratings transformed on the second rating scale, it holds

$$V_2(\tau(u)) = Var[\tau(U)] \quad (32)$$

which leads to

$$a_2 = a_1 = a. \quad (33)$$

As a result, the SOS hypothesis holds with the same SOS parameter a. The SOS parameter a is scale invariant when linearly transforming the user ratings in a mathematical way. However, it has to be clearly noted that subjective studies using different rating scales may lead to different SOS parameters. This has been observed e.g. for the results for speech QoE in [2].

As an implication, the numerical derivation (by solving the optimization problem) of the SOS parameter a for given MOS and SOS values can be done with linearly transformed user ratings, see Figure 2. Thus, the SOS parameter reflects the user rating diversity independent of the rating scale.
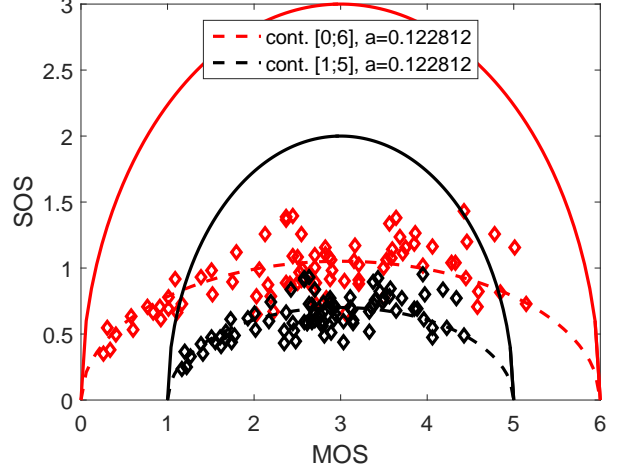

Figure 2. Transformation of user ratings from speech QoE results from rating scale $[0; 6]$ to rating scale $[1; 5]$ leads to the same SOS parameter a. However, the values of MOS and SOS (tuple depicted as diamond marker) as well as the maximum SOS for a given MOS (solid lines) are changing of course.

B. Computation of SOS Parameter

Due to the scale invariance of the SOS parameter, we consider now normalized user ratings $Z \in [0; 1]$ with $U^- = 0$ and $U^+ = 1$.

$$Z = \frac{U - U^-}{U^+ - U^-} \quad (34)$$

Then, the SOS hypothesis follows as

$$S(z) = \sqrt{a(-z^2 + z)} \quad (35)$$

for a normalized MOS value $z \in [0; 1]$ with the SOS parameter a.

For deriving the SOS parameter $a \in [0; 1]$ from the subjective normalized user ratings, for each test condition c the MOS u_c and the variance σ_c^2 are computed. Then, the least-squared error $L(a)$ between the subjective results and the SOS hypothesis is computed. The parameter a is derived by deriving the minimal least squared error.

$$L(a) = \sum_{c=1}^{R} \left(S(u_c)^2 - \sigma_c^2\right)^2 \quad (36)$$

$$\frac{d}{da}L(a) = 0 \quad (37)$$

$$\Rightarrow a = -\frac{\sum_{c=1}^{R}(u_c^2 - u_c)\sigma_c^2}{\sum_{c=1}^{R}(u_c^2 - u_c)^2} \quad (38)$$

C. SOS Parameter on a Binary Scale

We consider a binary scale where users accept ('1') or reject ('0') the observed quality. For any test condition c, we obtain a ration p_c of users accepting the service quality ($U = 1$). Hence, $1 - p_c$ reject the service quality ($U = 0$).

Thus, we observe the following mean and standard deviation for any test condition c.

$$E[U] = u_c = p_c \tag{39}$$

$$Var[U] = p_c(1 - p_c) \tag{40}$$

According to the SOS hypothesis and due to the scale invariance, we observe the following relationship between a MOS value u_c and ratio p_c for any binary scale.

$$S(u_c) \underset{\text{Eq.(35)}}{=} \sqrt{a(u_c^2 + u_c)} \underset{\text{Eq.(39)}}{=} \sqrt{a(-p_c^2 + p_c)} \tag{41}$$

$$S(u_c) = \sqrt{Var[U]} \underset{\text{Eq.(40)}}{=} \sqrt{-p_c^2 + p_c} \tag{42}$$

$$\Rightarrow a = 1 \tag{43}$$

Therefore, we conclude $a = 1$. Using a binary scale (e.g. to ask about acceptance) leads to maximum user rating diversity.

REFERENCES

[1] ETSI Technical Report ETR 250. *Transmission and Multiplexing (TM); Speech Communication Quality from Mouth to Ear for 3,1 kHz Handset Telephony across Networks*. European Telecommunications Standards Institute, July 1996.

[2] Tobias Hoßfeld, Poul E. Heegaard, Martín Varela, and Sebastian Möller. QoE Beyond the MOS: An In-Depth Look at QoE via Better Metrics and their Relation to MOS. *Quality and User Experience*, July 2016.

[3] Tobias Hoßfeld, Poul E. Heegaard, Martín Varela, and Sebastian Möller. Scripts for the Computation of QoE Metrics beyond the MOS, July 2016. https://github.com/hossfeld/QoE-Metrics

[4] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. SOS: The MOS is not enough! In *3rd International Workshop on Quality of Multimedia Experience (QoMEX 2011)*, Mechelen, Belgium, September 2011.

[5] ITU-T P.Sup3 – Suppl. 3 to ITU-T Series P Recommendations. *Models for Predicting Transmission Quality from Objective Measurements*. International Telecommunication Union, March 1993.

[6] Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Springer US, August 2000.