Mean Opinion Score (MOS) revisited: Methods and applications, limitations and alternatives

Robert C. Streijl · Stefan Winkler · David S. Hands

Abstract Mean Opinion Score (MOS) has become a very popular indicator of perceived media quality. While there is a clear benefit to such a "reference quality indicator" and its widespread acceptance, MOS is often applied without sufficient consideration of its scope or limitations. In this paper, we critically examine MOS and the various ways it is being used today. We highlight common issues with both subjective and objective MOS and discuss a variety of alternative approaches that have been proposed for media quality measurement.

Keywords Customer opinion · Psychophysics · Subjective MOS · Objective MOS · Quality monitoring · Fault isolation · Service level agreements

1. Introduction

Most media professionals would like to express quality in a way that is commonly understood and that facilitates comparisons across different algorithms, different organizations, and time. The International Telecommunication Union (ITU) has defined the opinion score as the "value on a predefined scale that a subject assigns to his opinion of the performance of a system" [1]. The Mean Opinion Score (MOS) is the average of these scores across subjects. MOS has emerged as the most popular descriptor of perceived media quality. It has had great success in the domain of speech quality, and consequently it has also been used for other modalities such as audio, images, video, and audiovisual content, and in numerous applications, from lab testing to in-service monitoring. MOS is not only used to express the results of subjective tests ("subjective MOS"), but also as the output of objective measurement algorithms, which provide an automated alternative to subjective tests (often referred to as objective or predicted MOS).

R. C. Streijl
AT&T, 99 Bedford Street, Boston, MA 02111, USA
e-mail: robert.streijl@att.com

S. Winkler (✉)
Advanced Digital Sciences Center (ADSC), University of Illinois, 1 Fusionopolis Way, Singapore 138632
e-mail: stefan.winkler@adsc.com.sg, URL: adsc.illinois.edu

D. S. Hands
Microsoft, 2 Waterhouse Square (140 Holborn), London, EC1N 2ST, UK
e-mail: dahands@microsoft.com

MOS is now the “de-facto” metric used to quantify perceived media quality. The 5-point MOS scale (excellent, good, fair, poor, bad) in particular is extremely popular. This has been very beneficial in terms of raising awareness for the importance of the perceptual aspect of media quality, and there is a clear benefit to a reference indicator of perceived quality and its widespread acceptance. However, MOS is often used without sufficient consideration of how the data has been obtained and the inherent limitations and restrictions imposed by the design of subjective tests or objective metrics. Various assumptions and preconceptions about MOS and its meaning prevail that are unfounded or incorrect, for example with respect to the accuracy, reliability, or applicability of MOS. Not always is it used for the right reasons or in the right context.

In view of these issues, we critically examine MOS measurements and the way they are used today. We identify common issues with MOS, as well as decisions and trade-offs that need to be made when designing subjective experiments or using objective quality models. Finally, we discuss alternative approaches that have been suggested by other researchers. In doing so, we hope to stimulate more research on this complex and inter-disciplinary topic. In addition, this paper will help users of subjective and objective MOS better understand and interpret these types of measurements. Even though some of the cited works may address a specific modality, we believe the issues with MOS discussed here are very similar across speech, audio, images, video, multimedia, or other modalities, so we do not explicitly separate them in this paper.

The paper is organized as follows. Section 2 reviews subjective testing methodologies, in particular aspects related to the design of such tests and the analysis of subjective MOS data. Section 3 reviews considerations for objective metrics and their output, in particular issues with model tuning and validation. Section 4 discusses a number of common applications of MOS, its limitations, and alternative approaches to MOS that have been proposed for media quality measurement. Section 5 provides a brief summary of the issues, conclusions as well as suggestions for future research on the topic.

2. Subjective MOS

Many aspects of MOS are determined by the choices made in the design of the subjective experiments; we provide an overview in this Section. A change in any of the design choices may well change MOS.

2.1. From Psychophysics to MOS

The complex and multi-dimensional nature of media perception [2] has resulted in the adoption of rather general, relatively open approaches to gathering quality assessment data. Current methods for obtaining subjective media quality ratings are quite different in process and rigor from traditional psychophysical methodologies. In the latter, subjects are asked to detect the presence of some signal (e.g., a tone or light), and from the answers a detection threshold can be obtained. These methods are useful for examining basic detection of simple stimuli, but fail to account for user strategies or confidence in a response. Signal detection theory [3] was designed to measure users’ confidence in responding to some signal as well as to accommodate various strategies employed by users. In measuring the quality of media signals, the just noticeable difference (JND) measure is the most closely related to signal detection theory; it is discussed in more detail later in Section 4.

Psychophysical methods are particularly useful for examining subjective thresholds for the presence of, or changes in, stimuli; in the context of picture quality they can be used to identify visible changes in the quality of test signals. However, for stimuli such as audio or video, whose spatial and temporal properties are complex and dynamic, traditional psychophysical methods do not appear well-suited. Indeed, the nature of psychophysical experiments requires very distinct tasks and test stimuli, and the use of natural media clips presents a major methodological challenge.

In media quality evaluations, subjects are providing a general opinion of quality for some time-varying signal. In contrast to psychophysical methods, standard methods used in subjective quality assessment (more on those below) fail to provide specific evidence for the contribution of different properties of a signal affecting quality ratings, but do offer a means for obtaining general ratings of highly complex media signals.

Although some form of subjective quality testing has been used since the early days of telephony (e.g., speech intelligibility) and television (e.g., line resolution), standardized testing procedures were not in place until the 1970s. In order to arrive at a general opinion of media quality, various standardized and non-standardized methodologies have been applied. Typically, the output from these tests is reported as a MOS value. Lewis [4] indicates that "MOS is a Likert-style questionnaire…" where a series of questions are asked, and respondents are required to provide answers on a single scale. Likert scales are also used in other disciplines, e.g. experimental psychology, human factors, and usability testing. Lewis further identifies criteria essential for robust media quality testing, such as reliability (measurement consistency), validity (measurement of the intended attribute), sensitivity (responding to specific experimental manipulations), number of scale steps, and factor analysis, which we will also explore in the following Sections.

2.2. *Subjective Test Design*

First and foremost, the application as well as the modality or modalities to be tested have to be decided. Single-modality tests (e.g. speech-only, or video-only) are generally easier to set up, and most subjective testing standards address single modalities.

The application further determines parameters such as frame rate and resolution for video or sampling rate and frequency range for audio. The range of impairments again affects the discriminating power of the test, as well as how the quality range in the test is mapped onto the rating scale by subjects.

Next, it is important to know what type of test is most suitable for the application at hand. ITU-R Rec. BT.500 [5] gives some guidelines in its Table 2, "Selection of Test Methods". Watson and Sasse [6] point out that multimedia conferencing requires different tests than any of those specified in ITU-T Rec. BT.500 [5] (broadcast television) or ITU-T Rec. P.920 [7] (audiovisual). They also indicate that not all multimedia components of a particular content type have to be tested similarly across different applications. For example, in multimedia conferencing applications it may make sense to focus on the audio rather than the video, as users concentrate more on the former.

Only certain test methods such as single stimulus continuous quality evaluation (SSCQE) are suitable for long clips. When interpreting data obtained from such studies, one must be aware of forgiveness [8] and recency [9] effects, where impairments that occurred some time ago tend to be "forgiven" or forgotten by the subject, while those occurring more recently tend to have a greater negative impact on the perception of video quality. Most other methods (see Section 2.3) require test clips of short duration (typically less than 15 seconds), for which memory effects are not so significant.

Another question that falls into this category is the general quality level of test clips. Double-stimulus methods or forced-choice methods are typically more suitable for small impairments in a test, whereas other methods such as single-stimulus can be used when subjects are presented with a wide impairment range.

The panel of subjects also has to be selected with care. Issues include:

- How many people should be used as evaluators? ITU-T Rec. P.911 [10] mentions "the possible number of subjects […] is from 6 to 40"; ITU-R Rec. BT.500 [5] recommends a minimum of 15. The number can be guided by the acceptable experimental error, as is done in medical experiments [11], or the desired confidence interval.
- Is the composition of the panel representative of the target audience for the application of interest? Demographics such as gender, age, ethnicity, education of the subjects are relevant and can affect MOS [12].

- Should expert or non-expert subjects be used [13,14]? Experts typically are in better agreement on the quality of a specific clip, so a smaller panel can suffice. At the same time, experts are more critical than non-expert subjects, which may skew ratings towards the lower end of the scale and fail to represent well the opinion of the intended target audience.

The last consideration is the testing environment. Each modality has its own specific issues; for example, for speech tests, the language spoken is an important factor that can affect the outcome of the test; for audio tests, the use of headphones or speakers has to be considered; for video, the type and properties of the display can have an impact on ratings [15].

ITU recommendations specify the test environment in detail, assuming a controlled lab setup. This may be considered "unnatural" when compared to the way a user would typically experience the content (e.g. watching TV in the living room, having a mobile phone conversation on the street). Subjects may be watching content in an experiment that they would not normally watch otherwise, or the test material may not be engaging to viewers because of the short duration of typical test clips [16]. Also, subjects are less tolerant of errors in short sequences, and video impairments appear less annoying to viewers consuming content in more natural environments [17].

2.3. Subjective Testing Methods

There is a large variety of subjective testing methods available, described in ITU recommendations [5,7,10,18-20] and other standards documents (see Table 1). Corriveau [21] provides a useful introduction to the various different methods. The main characteristics can be summarized as follows:

- Single-, double-, or multi-stimulus, i.e. the number of test clips to be compared in a single trial. Multiple stimuli may be presented simultaneously (e.g. side-by-side) or sequentially.
- Number of times a clip is presented to subjects (once, twice, or even multiple times).
- Presence of a reference clip, either explicitly (subjects know which one is the reference) or hidden.
- Subjects may rate the test clips only, both test and reference clips, or the difference between them.
- Depending on the interactivity of the voting process and setup of the test, one or more subjects may rate the clips in parallel.
- Ratings may be collected time-discretely (one rating per clip) or continuously (one rating per time interval). This also determines the length of test clips that can be evaluated.
- High and low anchor clips (either implicit or explicit) can help subjects "calibrate" their rating scale.

Subjects are typically given training before the actual test to familiarize them with the interface as well as the range of impairments in the test. The design and execution of this training phase can also have an impact on a subject's ratings. For example, central tendency bias (avoidance of the extreme rating categories) is a known problem in subjective experiments [22]. One of the purposes of the training phase is to encourage subjects to use the full range of the rating scale.

Researchers have compared various testing methods with one another and typically found high correlations between them. A comparison for many common testing methods, including absolute category rating (ACR), double-stimulus impairment scale (DSIS), double-stimulus continuous quality scale (DSCQS), and subjective assessment methodology for video quality (SAMVIQ), is done in [23] using mobile video as test clips. The authors of that study found very high correlations between MOS from the different methods (linear and rank-order correlation coefficients range from 96% to 99%), and no significant differences between methods. The same study also compares the assessment times for each method and found the following ranking (from fastest to slowest): ACR, DSIS, SAMVIQ, DSCQS. Finally, participants were asked to rate the ease of evaluation for each method, which resulted in the following ranking (from easiest to hardest): ACR-5, DSIS, SAMVIQ, DSCQS, ACR-11 (5 and 11 represent the number of discrete scale levels). It is interesting to note the large impact of scale on perceived ease for ACR.

SAMVIQ is also compared with ACR in [24], which addresses the differences in accuracy, granularity, and consistency between the two methods for both images and video. While SAMVIQ is found to require fewer subjects for the same MOS accuracy, it also takes more time in the rating process because of its interactive design.

Pinson and Wolf [25] compared both single-stimulus and double-stimulus continuous quality evaluation methodologies. They also proposed a way to convert time-varying continuous ratings (from SSCQE) into discrete ratings, although this can be tricky due to the effects of recency and forgiveness mentioned earlier.

2.4. *Subjective Testing Scales*

The most common standard rating scale used to derive MOS values is a category rating scale with five discrete levels. Due to the attachment of labels from "excellent" to "bad" to these levels, it is not only non-linear, i.e. the levels are not equi-distant across the scale (for example, the perceptual distance between "fair" and "poor" is larger than the distance between "poor" and "bad"), but also language-dependent (the levels and distances between them depend on the language in which the experiment is carried out) [26,27].

Despite the popularity of the 5-point MOS scale, there are other granularities, such as discrete scales with 7, 9, or 11 points, and (nearly) continuous scales. In theory, scales with higher granularity can result in smaller standard deviations of MOS [28]; in practice however, these differences turn out to be insignificant. For example, Huynh-Thu et al. [29] compared 5-, 9-, 11-point discrete and continuous scales for scoring video quality. They found no statistically significant differences between MOS or confidence intervals using the different scales. One of the reasons for this may be that people are able to reliably distinguish only a finite, limited number of levels of a certain quantity. In a classic paper, Miller [30] found this number to be 7±2 for many different types of experiments and stimuli. The potential higher resolution of continuous scales is lost in the noise due to the limits of human information processing capability.

It is also instructive to compare the MOS of viewer groups from different labs. We use data from the Video Quality Experts Group (VQEG)'s FRTV Phase I test [31] here, which were collected in 8 different labs from around the world. Even though correlations between the various labs are in the range of 0.9-0.95 for the most part, there are substantial variations, which were confirmed by an analysis of variance (ANOVA). In particular, viewers in different labs had quite differing opinions about the absolute quality range of the sequences (Figure 1); for example, a clip with an average rating of 30 in one lab might score 60 in another [32].

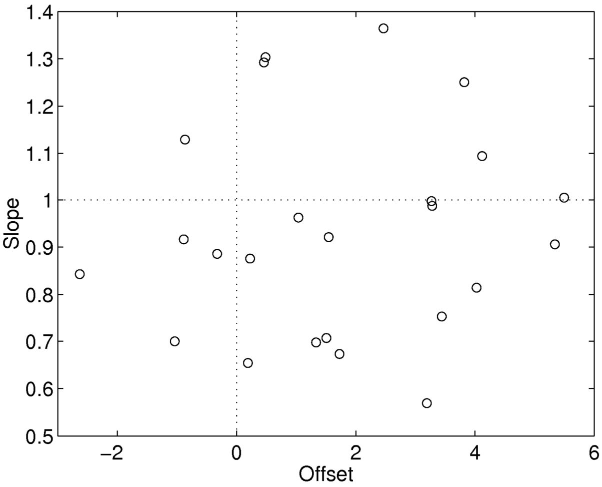

Figure 1: Comparing Difference MOS (DMOS) from 8 different subjective testing labs [32]. Every data point represents slope and offset parameters of a linear regression line between a pair of labs. Rating scale is 0-100. The slope in particular deviates quite far from 1, indicating differing viewer opinions about quality range in different labs.

2.5. *Subjective Test Results*

After a subjective test has been conducted, the individual ratings are averaged to obtain the MOS for each clip.

Before doing this, the raw data are normally screened for outlier subjects. A typical screening procedure is defined in ITU-R Rec. BT.500 [5]. The screening can be useful to detect and remove unreliable subjects from the data; at the same time, it may eliminate opinions that are just as valid as everybody else's. After screening, the MOS inevitably represents the view of a majority.

It is also important to remember that MOS is not a precise measurement – it is a statistical quantity (something that is often ignored when comparing it to objective measurements, as will be discussed further in Section 3). MOS is simply the average of a distribution of a finite number of individual ratings, and as such it is one statistic among many (standard deviation, confidence interval, kurtosis, etc.) describing that distribution. This is also one of the criticisms raised in [33], where it is shown that MOS implicitly assumes homogeneity among subjects due to the arithmetic averaging. Finally, the analysis of MOS data should adhere to statistical principles. For example, selection of parametric or non-parametric statistical tests should be appropriate to the type of data under investigation.

In summary, MOS is not just your average noisy measurement; the subjectivity and other factors discussed in this Section add to its complexity. What this means for objective MOS is the topic of the next Section.

Table 1: Overview of ITU Recommendations on subjective and objective quality measurement of speech, audio, video, and audio-visual signals.

Modality	Subjective	Objective
Speech (non-conversational)	ITU-T Rec. P.800 (1996) ITU-T Rec. P.806 (2014) ITU-T Rec. P.830 (1996) ITU-T Rec. P.835 (2003)	ITU-T Rec. P.563 (2004) ITU-T Rec. P.564 (2007) ITU-T Rec. P.862 (2001) ITU-T Rec. P.863 (2014)
Speech (conversational)	ITU-T Rec. P.800 (2008) ITU-T Rec. P.805 (2007) ITU-T Rec. P.1302 (2014)	ITU-T Rec. G.107 (2014) ITU-T Rec. P.561 (2002) ITU-T Rec. P.562 (2004)
Audio	ITU-R Rec. BS.1116 (2014) ITU-R Rec. BS.1284 (2003) ITU-R Rec. BS.1534 (2014) ITU-R Rec. BS.1679 (2004) ITU-T Rec. P.830 (1996) ITU-T Rec. P.913 (2014)	ITU-R Rec. BS.1387 (2001)
Video	ITU-R Rec. BT.500 (2012) ITU-R Rec. BT.1663 (2003) ITU-R Rec. BT.1788 (2007) ITU-R Rec. BT.2021 (2012) ITU-T Rec. J.140 (1998) ITU-T Rec. J.245 (2008) ITU-T Rec. P.910 (2008) ITU-T Rec. P.913 (2014)	ITU-R Rec. BT.1683 (2004) ITU-R Rec. BT.1866 (2010) ITU-R Rec. BT.1867 (2010) ITU-R Rec. BT.1885 (2011) ITU-R Rec. BT.1907 (2012) ITU-R Rec. BT.1908 (2012) ITU-T Rec. J.144 (2004) ITU-T Rec. J.246 (2008) ITU-T Rec. J.247 (2008) ITU-T Rec. J.249 (2010) ITU-T Rec. J.341 (2011) ITU-T Rec. J.342 (2011) ITU-T Rec. P.1202 (2012)
Audio-visual (non-conversational)	ITU-T Rec. P.911 (1998) ITU-T Rec. P.913 (2014)	ITU-T Rec. P.1201 (2012)
Audio-visual (conversational)	ITU-T Rec. P.920 (2000) ITU-T Rec. P.1301 (2012)	ITU-T Rec. G.1070 (2012)

	ITU-T Rec. P.1302 (2014) ITU-T Rec. G.1091 (2014)	

3. Objective MOS

Although subjective quality tests are essential to understanding customer opinions of media, they are costly and time-consuming to perform, and any single test is limited in scope. As a result, objective quality models are an important alternative to subjective tests, particularly for media quality benchtesting and in-service monitoring, where subjective tests are impractical. Various types of objective models are known, including arithmetic models such as peak signal-to-noise ratio (PSNR) or mean squared error (MSE), statistical models such as structural similarity (SSIM), parametric network planning models (e.g., ITU-T Rec. G.107 [34], G.1070 [35]) and perceptual models (see [36] and Table 1 for an overview of standards activities, and [37-39] for recent surveys of quality metrics). This discussion is limited to perceptual models, because they are designed to emulate subjective quality ratings, and as such output a predicted MOS.

Naturally, there is a close relationship between subjective assessment and objective measurement. Objective models are trained and tested against media for which subjective scores are available. Consequently, all the considerations discussed in the previous Section affect objective MOS as well. The relationship is established through a process of tuning (a.k.a. calibration or training), where the model outputs are aligned with subjective MOS as closely as possible. Once a model has been trained, it is validated (a.k.a. testing or benchmarking), usually again with subjective MOS data. Tuning and validation are not only the most common uses of (subjective) MOS, but also the most suitable ones (cf. Table 2). A good overview of various methods along with practical guidelines for the calibration, validation, and comparison of objective models can be found in [40].

Table 2: Summary of MOS suitability, limitations, and alternatives for different applications.

Application	**MOS Suitability**	**MOS Limitations**	**Alternatives to MOS**
Metric Tuning & Validation (Section 3)	Good	Positive-only test Scalar value	Stress testing Failure characteristics
Quality Monitoring (Section 4.1)	Good	Thresholds	Acceptability JND
Fault Isolation (Section 4.2)	Poor	Global measure	Multi-dimensional quality
Service Level Agreements (Section 4.3)	Poor	Short-term quality	One impairment per time frame Mean time between failures Long-term quality tests

3.1. Tuning / Calibration

As with other measurement devices, there is a need to calibrate objective quality assessment algorithms in order to obtain meaningful and reliable results. Calibration is a comparison between measurements – one of known magnitude or correctness made or set using one method (also known as a reference) and another measurement made using a similar, second method. In the case of objective models, this is done by comparing objective MOS predictions with subjective MOS, ideally for a large database of media. The aim is to match both as closely as possible. This has a number of implications for the meaning and usability of objective MOS.

One important issue here is that the scale of objective MOS is generally different from subjective MOS. The scale of subjective MOS is determined primarily by the range of content quality and impairments present in a given subjective test. The limits of the scale range are set through training of the subjects using anchor clips. The scale of objective MOS on the other hand can in principle be infinite.

The MOS tuning process therefore focuses on making sure the trends and relationships between subjective MOS and the MOS predicted by an objective model match well. This is typically done by evaluating correlations between those two datasets. High correlations mean a good match has been achieved (while there is no generally accepted level for what constitutes a sufficiently high correlation, one common benchmark is the correlation of PSNR and MOS for the same dataset [31]). However, correlations are independent of range and scale and do not give any indication about how well subjective and objective MOS values correspond in absolute terms.

As a result, fitting functions are often used to not only maximize the linear correlation, but also to match the range of objective and subjective MOS scales. Both linear and non-linear (e.g. polynomial or logistic) functions are commonly applied in this process, the main argument for the latter being that non-linear functions can compensate for any saturation of subjective MOS towards the ends of the scale, an effect that objective MOS would not necessarily exhibit. This fitting allows the computation of the residual prediction error of a model, which is another common performance criterion.

A question remains whether the fitting function should be considered part of the model (in other words, generic), or part of the data (i.e. specific to a given set of subjective tests). As an example, ITU specified a mapping function for transforming the raw result scores of its method for perceptual evaluation of speech quality (PESQ) to a linearized MOS in a separate recommendation [41].

To summarize, it is unlikely that a specific model would be able to predict subjective MOS values for any given subjective experiment without some adjustment of the objective MOS scale via a fitting function. In practice, this type of calibration is generally required for any model in an application where it has not been used or tested before.

3.2. Validation / Testing

Although recent novel methods for objective quality assessment [42,43] have been able to reduce or eliminate the need for using subjective MOS in training (but not validation of course), validation and testing of objective models is still an essential part of the work. Validation of objective models, such as [44] or the tests performed by VQEG, is based on a comparison of objective model outputs with subjective quality scores for a set of media clips. VQEG validation tests are used by the ITU to produce new standardized objective measurement methods (cf. Table 1); an alternative streamlined approach to validating objective models has also been proposed [45].

An important issue is that – as mentioned earlier – subjective MOS is not a precise number, but a statistical measurement. As long as the predicted MOS lies within the confidence interval of the subjective MOS value, it can be considered correct. This means that not only any kind of performance evaluation criterion (correlation, residual error, etc.) would have to come with a confidence interval, but also objective MOS itself, yet model performance or objective measurements are rarely reported in this way.

For MOS to be used successfully in this scenario, it cannot be treated as a simple scalar value. Often we rely on the assumption that there should be some homogeneity or correlation in judgments from different subjects, but that is not at all clear [33,46]. It is true to some extent if the content is limited to a single type of distortion; a slight broadening of the distortion characteristics results in much more complexity and thus disagreement between subjects. Therefore, it is important to take the disagreement and sometimes conflicting opinions of subjects into account.

Traditional linear regression or root mean squared error (RMSE) calculations make additional problematic assumptions: they assume that residuals have a Gaussian distribution and equal variance. Because of the discrete and limited MOS rating scales as well as human nature, this is generally not the case for subjective quality ratings [28]; in fact, they follow an ordered multinomial distribution [47]. Therefore, it is preferable to adopt approaches that do not rely on these assumptions. A simple way of achieving this is using weights that are equal to the reciprocal of the variance of each measurement (MOS) for each test clip; this is also referred to as variance-weighted regression, a special case of generalized least squares [47]. Note however that the accuracy of this method depends on the accuracy of the individual variance estimates [48].

More advanced statistical tools such as the Generalizing Linear Model (GLZ) can provide an additional benefit: they do not assume an equi-distant rating scale, which is another flawed assumption for most MOS scales (see Section 2.4). The GLZ model yields a probability for each possible rating category, which is much more insightful than MOS [47].

Another such approach addressing this issue [49] proposes a rank agreement measure (RAM) from a geometric representation of subjects' rank order preferences. The RAM can be calculated for a subjective or an objective rank. The approach is based on the premise that the qualification of objective algorithms should depend on the level of agreement between subjects. They identify the mean Spearman rank as a useful RAM and suggest that an algorithm for objective quality prediction can be approved if it is better than the mean RAM of subjects.

Wu, Hu, and Gao [50] represent the visual quality of images as a distribution rather than a single scalar value. They propose a structural regression algorithm to cope with learning and predicting this data. Furthermore, they introduce a reliability-sensitive learning method, which weights quality ratings by their reliability (essentially the number of ratings per sample), as well as a refinement strategy that iteratively improves samples with lower reliability by propagating information from samples with higher reliability.

Completely different methodologies to validating objective metrics have been proposed recently, using approaches akin to software testing, in that they aim to expose "bugs." i.e. errors, vulnerabilities, and failure characteristics, of a quality metric, rather than to demonstrate that it satisfies certain specifications [51-53]. Reibman [54] generalizes these approaches to define subjective tests with a high likelihood of exposing misclassification errors of objective quality metrics, whereas conventional subjective tests generate such samples only randomly. These misclassifications can be categorized as false ranking, false differentiation, or false tie. The algorithm proposed is able to determine the best images for a pairwise subjective test for this purpose using a set of existing objective metrics. This is particularly useful for comparing the accuracy of objective metrics across distortion types or reference samples.

Other examples of such approaches include checking the output for content with the same PSNR, distinguishing un-degraded from heavily impaired images, simple transformations such as cropping, monotonically increasing distortion levels, etc. [53]. One of the advantages of these checks is that they require little or no subjective testing. Unfortunately, such tests are severely under-utilized in objective model validation; most still rely exclusively on traditional (positive-only) regression approaches.

4. MOS Applications

In this Section, we discuss various practical applications and uses of MOS as well as its limitations and suggest possible alternatives. We have grouped them into three areas, based on the main use cases that have emerged from numerous discussions with media professionals and actual users of quality of service (QoS) and/or quality of experience (QoE) tools, for example codec designers, service providers, telecom operators, and other engineers. The application areas we consider here are quality monitoring, fault isolation, and service level agreements (cf. Table 2). The nomenclature may be service provider oriented, but the applications are easily translated to other use cases.

4.1. Quality Monitoring and Alerts

The purpose of quality monitoring and alerts is to notify users (e.g. operations and support teams) of potential problems. For these notifications to be useful, they need to be near-instantaneous, i.e. on the order of a few seconds. For an operator, the goal of monitoring is to pre-empt customer helpdesk calls, as those increase operational costs.

MOS is relatively well suited for this type of application (at least in theory), because (a) MOS represents the overall quality and captures various possible problems, and (b) a MOS value can be produced for short segments of content (e.g. a few seconds), which makes it possible to capture short-term quality fluctuations and keeps delays low.

However, for any given scenario in which MOS is reported, it is unclear what threshold values should be applied to identify problems or determine acceptability. Which MOS value is good, or good enough? As an example, [55] suggests values between 3 and 4 for IPTV applications. But can this threshold be static, i.e. a universal single value for all models and applications? For example, it may be argued that the quality of high-definition (HD) content with a MOS of 4 is not the same as standard-definition (SD) content with the same MOS. The subjective scores would be modified by expectations, and objective models would take this into account through tuning. In other words, the same MOS may actually mean something different in terms of absolute quality.

Unfortunately, service providers often have conflicting expectations: MOS values should scale independently of the resolution (e.g. both unimpaired SD and HD video should have a MOS of 4.5 or above), while MOS for HD should generally be higher than MOS for SD, which obviously cannot be achieved with a single MOS scaling or threshold.

Differences in MOS thresholds can be due to two things:

1. MOS scales are dependent on the range of qualities present in a given subjective test. Speech quality is quite well-defined, and therefore its range is easier to cover; for video on the other hand this is more complex. In a hypothetical experiment that includes both SDTV and HDTV content, for example, all HDTV content may be rated above 4, for example, whereas in an HDTV-only experiment, it is likely to cover a wider range. Similar things apply to content type (e.g. sports vs. cartoon), bitrates, or any other test parameters. Choosing a suitable MOS threshold therefore means having to match the scale used in the original experiment(s) with the actual content being monitored.

2. There is also a commercial aspect to the selection of a MOS threshold. Service providers have different priorities for different services and may decide to set very different quality (bandwidth) targets for each service offering (channels or programs). A high-value subscription-based sports transmission or prime-time movie will typically be shown at higher quality than a morning cartoon or late-night reruns. Different MOS thresholds would have to be applied in these cases.

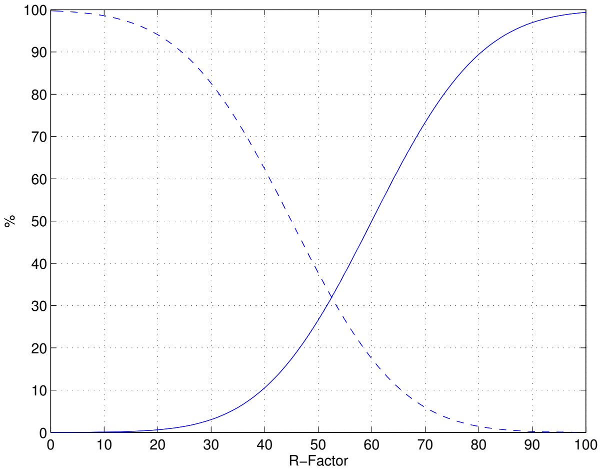

Figure 2: Service acceptability mapping functions. The percentage of users experiencing “good or better” (solid curve) and “poor or worse” (dashed curve) service are obtained from the R-Factor, an indicator of voice transmission quality [34].

Another problem is comparing MOS results across different studies or results from objective models that are paired with different subjective tests, which should be avoided and discouraged. Different studies use different equipment, type of content, quality of source content and impairments of that content, etc. [16,56]. ATIS 0800008 [56] describes such parameters that need to be reported as part of a subjective tests and objective models; ITU-T Rec. P.800.2 [57] serves a similar purpose.

One of the issues is that MOS measures the amount of satisfaction rather than the acceptability or acceptance of a service. In voice quality, "Good or Better" (GoB) has been used as a basis for acceptability [34]; in a perhaps overly simplistic manner, it is derived directly from a mapping of MOS values (see Figure 2). De Koning et al. [58] find that video requires a different (and possibly content-dependent) mapping function. Jumisko-Pyykkö et al. [59] even recommend a bi-dimensional approach to quality evaluation for this purpose, by considering satisfaction and acceptance separately.

Another useful alternative here may be one that does not focus on quality but on noticeable impairments. One such technique is based on just noticeable difference (JND), as mentioned briefly in Section 2.1. The JND can be determined for different modalities and any kind of impairment. JND is the impairment level at which test subjects detect a difference between two stimuli some proportion of the time. Depending on the psychophysical method used, the difference is typically considered noticeable when subjects report it in 50% or 75% of the tests (see Figure 3). This "just noticeable" impairment level is defined as 1 JND; a higher number of JNDs means higher impairment levels and a poorer quality [60]. JND has been defined as a subjective method and has been used extensively in psychophysical studies. There are also objective models of JND [61]. The advantage of JND measurements is that – contrary to MOS – they are in absolute and statistically meaningful units and are less likely to be subject to context effects [60].

JND is very reliable at or near the perceptual threshold, which makes it useful for applications operating in this range (e.g. encoder tuning). However, it becomes less intuitive for larger quality differences, at which many commercial services operate. Multiple units of JND can still be related to a certain likelihood or percentage of observers detecting a difference, but they are not easy to interpret or use in practice. Maximum likelihood difference scaling (MLDS) was proposed as an experimental methodology and model to map difference measurements of the kind "difference between A and B is greater/smaller than the difference between C and D" to a supra-threshold difference scale [62], which has been successfully applied to image compression.

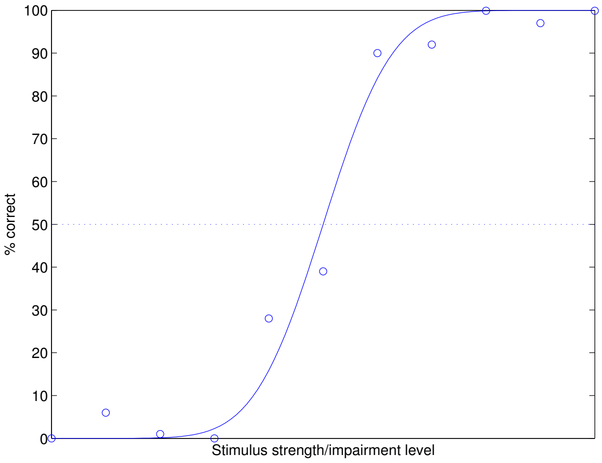

Figure 3: Typical outcome of an experiment for establishing the just-noticeable difference (JND) level (dashed line, at 50% of correct responses). Circles: empirical measurements; curve: fitted psychometric function.

4.2. Fault Isolation

It is generally not enough to establish the existence of a problem; it is necessary to identify the root cause in order to fix it. Unfortunately, MOS tells users little about the causes of a problem. Without additional measurements or detailed information, fault isolation or troubleshooting is nearly impossible.

Therefore it is important to consider the multi-dimensional aspect of media quality and explicitly measure and report the different dimensions. Virtanen et al. [63] indicate that quality is not a "single monotone dimension". Sen [40] speaks of dimensions in a "quality space". Preminger and Van Tasell [64] indicate that speech quality is of a multi-dimensional nature – with uni-dimensional methods the measurements are essentially judgments, where one or several of the individual quality dimensions may influence the subject's preference. A number of studies emphasize the need to find the quality variables of each dimension and to understand the relationship and weights of each of these dimensions

[27,46,65]. Although there has been quite a bit of research into the multiple dimensions of quality, the use of multi-dimensional models for quality assessment for the various modalities is still relatively immature.

4.2.1.Multi-dimensional Quality Perception

To determine the perceptual dimensions of quality, two approaches are commonly used:

1. Scaling perceptual differences of pairwise stimuli, which are then mapped into a multidimensional space using multi-dimensional scaling (MDS) procedures [66,67].
2. Asking subjects to rate stimuli on a set of bipolar scales (e.g. warm-cold) according to the semantic differential (SD) [68], and reducing the dimensionality by means of factor analysis.

An important difference between these two approaches is that SD presents predefined dimensions and scales to subjects, whereas perceptual difference scaling with MDS does not. Therefore, the set of attribute scales for SD has to be chosen with care, often through pilot tests, and their number can be rather large (see below). SD results are usually easy to interpret; however, if a certain dimension is not part of the pre-defined set, this information is lost. MDS on the other hand can reveal "hidden" quality dimensions, but can be harder to interpret; also, a full pairwise test is often not practical because of the large number of possible stimulus pairs [69].

For semantic differential, it might be reasonable to assume that the most apparent criterion for a user is a good-bad dimension. However, for the purpose of finding perpendicular dimensions reflecting different features that together form the integral quality judgment, it can actually be counterproductive to consider a separate dimension reflecting integral quality itself [70].

Note that MOS methods and scales can still be used to evaluate individual quality dimensions; in other words, appropriately instructed subjects can be asked to rate a specific quality dimension on a MOS scale (see Section 2.4), and the averaged responses across subjects then represent MOS for that dimension.

Ghinea and Thomas [71] define Quality of Perception (*QoP*), which includes a user's satisfaction with multimedia clips as well as a user's ability to understand, synthesize and analyze the information content of such presentations. This may be considered a two dimensional modeling of video quality.

In an effort to minimize the sampling error associated with individual differences in subject's taste or preference, Voiers [65] designed the Diagnostics Acceptability Measure (DAM). DAM is a variation of the SD technique and uses different attributes of a signal. DAM is a subjective measure of quality, based on the ability of a group of listeners to detect different types of distortions. It combines both direct (isometric) and indirect (parametric) approaches to acceptability evaluation. As a result, DAM is time consuming and requires trained listeners. DAM scores are also used by Sen [46] in a principal component analysis (PCA) and MDS to show the multi-dimensional nature of speech.

Wältermann [70] identifies quality dimensions of speech using MDS. Starting from 13 and 28 antonyms and corresponding bipolar scales (see Figure 4), he finds three common dimensions, namely discontinuity, noisiness, and coloration, for both narrowband and wideband speech transmission, with an additional fourth dimension (high-frequency distortion) specific to the wideband case. These dimensions capture a wide variety of conditions/distortions; furthermore, many distortions can intuitively be assigned to one or more specific dimensions (e.g. packet losses to discontinuity). A simple linear model is able to capture the relationship between perceptual dimensions and integral quality quite well.

Interrupted								Continuous
Distant								Close
Crackling								Not crackling
Thin								Thick
Not noisy								Noisy
Muffled								Not muffled

Shaky								Steady
Indirect								Direct
Dark								Bright
Unintelligible								Intelligible
Not hissing								Hissing
Clear								Unclear
Distorted								Undistorted

Figure 4: Semantic differential (SD) scales and corresponding antonyms used in a narrowband speech experiment [70].

The findings clearly depend on the modality and the distortions present in the experiment. A review of 9 studies of quality dimensions of synthetic speech concluded that the 5 universal quality dimensions were naturalness, prosodic quality, intelligibility, disturbances, and calmness [69]. Watson and Sasse [6] use quality dimensions to find an appropriate scale. An example quality dimension for multimedia conferencing speech is “choppiness”, which uses terms such as ‘broken’, ‘cut up’ and ‘irregular’. In video, many perceptual dimensions have been defined; for example, block distortion, blurring, jerkiness, etc. [72].

Especially for MDS-based approaches, it can be useful to combine conventional quantitative psycho-perceptual evaluation with a descriptive qualitative evaluation based on the individual’s own vocabulary, which can make it easier to interpret the resulting quality dimensions. Interpretation-based quality (IBQ) [73] and open profiling of quality (OPQ) [74] are good examples of such approaches.

Egger et al. [75] on the other hand combine semantic differential scores for a variety of attributes and MOS scales for different questions in a test on interactive video communication; by means of PCA, they find only a weak relation between subject responses on those two scales and highlight certain limitations of SD in this scenario.

4.2.2.Multi-dimensional Quality Measurement

Perceived (subjective) and measured (objective) dimensions of quality may be different, because they have somewhat different purposes. Perceived dimensions tell us about the criteria people use to rate media quality. Measured criteria are there to break quality down into various components (e.g. impairments) from a measurement perspective [76], but more importantly perhaps, they should also reflect the various faults and impairment sources we are trying to isolate.

In practice, many objective quality models metrics compute and/or analyze a multitude of features and/or artifacts before coming up with an overall estimate of quality. This is especially true for those that do not rely on a reference – see e.g. [77] for a review of no-reference image and video quality metrics. They often rely on a hierarchical analysis of measurements, as shown in Figure 5: various network- or media-level parameters and measurements form the basis for Key Performance Indicators (KPIs). KPIs are aggregated into Key Quality Indicators (KQIs), which in turn form the basis for Customer Experience Indicators (CEIs). While numerous individual KPIs, KQIs, and CEIs have been defined by various standards organizations, the challenge is establishing reliable integration functions or mappings from one level to the next.

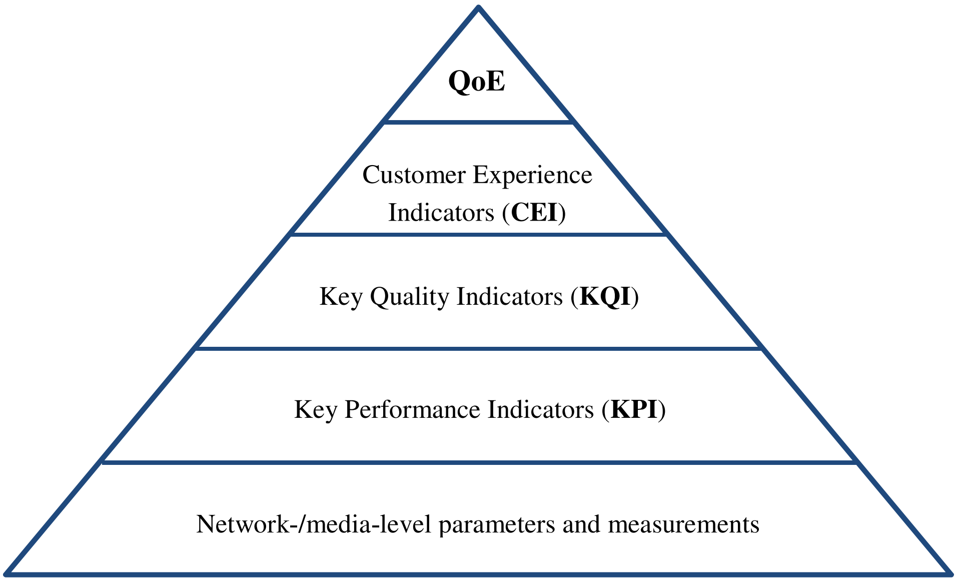
QoE
Customer Experience Indicators (CEI)
Key Quality Indicators (KQI)
Key Performance Indicators (KPI)
Network-/media-level parameters and measurements

Figure 5: Hierarchy of quality assessment indicators.

In this respect, the issue of fault isolation has also been addressed by the networking community [78], using multi-resolution data analysis and statistical data mining techniques to identify problems and diagnose their root causes in large video delivery networks [79]. However, these approaches have two shortcomings. First, they generally require instrumenting the system at all points and components along the distribution chain, which can be prohibitive in practice. Second, they have mostly focused on network and device performance, and have given only very limited consideration to media quality.

The importance of multidimensional analysis for quality assessment is also highlighted in a study by Keimel et al. [80]. Their approach treats the human visual system as a black box and derives quality from a number of features, without making any assumptions on their relationships. A similar approach is presented in [81] by constructing a k-dimensional Euclidean QoE space, where each dimension represents a parameter that may impact video quality. In a subjective test for a given set of parameters and values, “reference points” are obtained, each with a MOS. These allow inferring the MOS for a new set of parameters using a simple, lightweight search algorithm. Another study formulates multidimensional video quality assessment as a transformation between a multidimensional feature space and a quality space, where each space itself may be multi-dimensional [82].

The research on the multi-dimensional nature of media quality has recently been picked up by standards bodies as well. ITU-T Study Group 12 is studying multi-dimensional speech quality analysis from both subjective and objective perspectives (referred to as work items P.AMD, P.MULTI, and P.TCA). Likewise, VQEG is working on a project on the monitoring of audio-visual quality by key indicators (MOAVI for short) to understand and measure the effects of different impairments on quality.

4.3. Service Level Agreements

A Service Level Agreement (SLA) defines the level of service delivered to a customer in terms of contractual metrics; often, penalties can be imposed in case of non-compliance. A service provider (SP) may be interested in using MOS to offer/maintain quality-based SLAs with content providers and/or customers in order to differentiate pricing plans based on different levels of quality. They may also want to compare the quality of their service with the competition. For this type of application, users typically look at long-term quality behavior and trends, on the scale of a single program all the way up to weekly or monthly trends.

The Tele Management (TM) Forum has published a comprehensive guidebook on service level agreements [83]. It has also defined a hierarchy among KPIs, KQIs, and SLAs, asserting that SLAs can be defined in terms of product KQIs (or CEIs). Based on this hierarchy, Seo et al. [84] propose an architecture for defining and managing SLAs as well as detecting violations.

As Sullivan et al. [16] point out, designing and maintaining an infrastructure that aims for a service that is constantly scoring a maximum value is too costly and does not provide sufficient noticeable benefits. Unfortunately, one major problem with MOS is that little is known on how much and what type(s) of variations are tolerable. A typical operational/engineering approach is the use of "one-impairment-per-time-frame." e.g. 1 hour or 4 hours [85,86]. However, there are a number of issues with this approach, such as:

- Do customers evaluate their service or content in such long, defined time periods?
- How many error-free and non-error free intervals should be considered as part of these longer observation periods? Setting a fixed value would have to be paired with appropriate statistical constraints.
- Do users rate noticeable impairments similarly, independent of time of day, type of content, level of interest, etc.?
- Will users rate impaired content differently when it impacts a critical moment in a program as opposed to a non-critical one?

Some actual figures about the number of artifacts perceived by consumers and their acceptability are found in [87].

The Alliance for Telecommunications Industry Solutions (ATIS) IPTV Interoperability Forum (IIF) has identified desirable features that should be considered for in-service perceptual quality measurements, including frequency of impairments; duration of each impairment; how bad the impairment is (content sensitive perspective); what impairment it is. With these features and the previously mentioned issues, ATIS-0800041 [55] informally suggests a two-stage process for defining threshold "severities" for IPTV applications. The first stage involves defining the severity level of each error, depending on its duration and extent. The second stage aggregates the frequency and severity levels of multiple errors into an overall trouble level.

The ITU has been building an arsenal of subjective methodologies for specific applications over the last several decades. However, they are all based on controlled settings in a lab environment. Sullivan et al. [16] claim that testing in isolation according to the ITU standards is "appropriate for situations where reliability and small error variance are important" and that "when looking at electrical components, psychophysical approaches work perfectly well". But none of these methods are practical for determining the quality that has to be maintained in a service provider network. They argue that a human factors approach is more appropriate in situations where SPs are planning for how much bandwidth is needed per second to obtain an acceptable level of video quality, as well as how video quality compares among different SPs.

Staelens et al. [17] modified some of the standardized subjective tests by replacing short duration test clips with long naturally occurring high value (high motion) content, a living room environment as opposed to a controlled lab environment, and a more granular scale to better measure quality differences at the higher end of the scale. Specifically, they compared responses from subjects to frame freezes and blockiness artifacts from long tests with entire movies to those in traditional short tests. However, they eventually translate the results back to the popular 5-point scale.

One interesting alternative was proposed by Suresh and Jayant [88]. They describe an "intuitive, global" subjective metric called mean time between failures (MTBF) that represents how often visual artifacts are observed by a typical viewer. In addition, an "instantaneous" metric called 'Probability of Failure' (PFAIL) is introduced, which reflects the fraction of viewers that find a given video portion to be within acceptable quality levels. MTBF links artifact occurrence frequency and duration to video quality; if an artifact is persistent, the viewer simply continues to push a button, which gives an indication of how bad a visual impairment is. This method does not differentiate between "amounts" of impairments (small or large visible artifact), the location of the impairment with respect to the perceived content (center, edge/corner), or types of impairments. MTBF is calculated over a number of sequences as the inverse of the average PFAIL.

4.4. *Other Considerations*

When monitoring and fault isolation methods are in place, it seems natural to use MOS measurements as the basis for any adjustments to the system in order to optimize media quality, i.e. using MOS within a feedback loop. From the objective perspective, the topic of quality control of multistage systems has been studied in industrial process engineering [89]; media distribution can be considered a special case of such systems. The problem here is similar to the application of MOS for fault isolation (except the time scale is different). Without understanding the various quality dimensions and impairments, it is hard (albeit not impossible) to tweak the right parameters for improving overall quality.

From a service provider's perspective, "maximizing QoE" may have different objectives; these could be maximizing overall QoE for multiple users in a network, maximizing the QoE of certain individual users or groups, maximizing the number of "satisfied" users, etc. In this respect, overall MOS can lead to unfairness among customers, depending on how it is used [33]. In practice, instead of considering average QoE, service providers are more interested in customers with low perceived quality, who are at risk of switching. If monitoring (see Section 4.1) or optimization methods rely on MOS, the result may be over-provisioning, because users who already receive good quality are given even better service merely as a side-effect of improving the quality for those with the worst QoE. To remedy this problem and improve QoE management, Xu et al. [33] propose a MOS based on utility functions, which have been used for rate control in networking research.

Finally, how broad and encompassing can or should MOS be? Knoche et al. [90] highlight several properties of MOS-based approaches that may prove problematic when used in certain applications. They indicate that current subjective methods do not register some aspects of the subject's audiovisual system, such as unconscious effects; differences in judging, mood, and a-priori estimates. In addition, they claim that MOS allows for ambiguous results, which leads to complications in using MOS. To overcome these pitfalls, they introduced Task Performance Measures (TPM), which is a set of tasks such as repetition, memorization etc. that subjects perform; these tasks can be measured objectively. One example they offer is that they can measure the degradation in performance by wrong answers such as those introduced by McGurk effects [91], whereas a MOS score might miss this degradation.

Mullin et al. [27] indicate that standardized subjective methods are "cognitively mediated" and claim that other variables can influence the user's assessment of quality. They use a traditional Human Computer Interaction (HCI) evaluation framework that considers task performance, user satisfaction, and user cost. Likewise, Laghari et al. [92] present a comprehensive framework that considers factors from the human domain (demographics, expectations, role), contextual domain (e.g. social context, device, environment), technical domain (e.g. design, features), and business domain (e.g. brand, pricing). When quality or QoE is defined in such broad terms, it becomes more and more difficult to describe them with a single number such as MOS. Similar issues around QoE are explored in [93].

5. Summary and Conclusions

MOS is a wide-spread and popular measure of media quality. Large numbers of subjective MOS data [94] and objective quality metrics [37-39] have been made available. At the same time, MOS values are prone to misuse or misinterpretation. Choices made in the design of subjective experiments on media quality have an important influence on MOS values and need to be taken into consideration when analyzing and using MOS data. Objective media quality metrics rely on data from those subjective experiments for tuning and validation, and are therefore affected by the same choices and factors.

We have presented a number of important methods and practical applications of MOS, its various respective limitations, and some alternative approaches that have been proposed to overcome those (cf. Table 2). Many of these alternatives are supported by solid theoretical foundations and numerous studies; however, their practical application is still lacking. For example, while multidimensional quality is a well-known concept, with a multitude of subjective experiments and available mathematical tools, complemented by an equally large number of objective models that rely on quantifying various distortions, it is still nearly impossible to link some of these quality dimensions to actual faults or root causes in a complex system.

An essential task for the media quality research community is to provide relevant guidance for users of MOS to enable them to make meaningful measurements and interpret the results correctly. Educating users of MOS about its limitations is particularly important. Standards organizations also have to play their part here. Even in the research community, further education efforts are necessary to overcome issues such as the tendency to rely on simplistic validation methods of objective models that are based on problematic assumptions.

Finally, we believe that more research efforts need to be directed towards addressing practical applications and system issues, i.e. answering questions such as how quality measurements can be used to trigger alarms, how those alarms can help identify and fix quality problems, or how impairments should be aggregated over longer pieces of content. Alternative subjective and objective quality indicators that are able to address some of the shortcomings of MOS can put a more versatile set of quality measurement tools at users' disposal, from which they can then pick the most appropriate ones. Ultimately, what matters is not MOS, but managing QoE.

Acknowledgments S. Winkler is supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR).

References

1. ITU-T Rec. P.10 (2006) Vocabulary for performance and quality of service.
2. Coren S, Ward LM, Enns JT (2003) *Sensation and Perception.* 6th edition, Wiley.
3. Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics*. London: Wiley.
4. Lewis JR (2001) Psychometric properties of the mean opinion scale. *Proc. HCI International,* vol. 1, pp. 149-153, New Orleans, LA.
5. ITU-R Rec. BT.500-13 (2012) Methodology for the subjective assessment of the quality of television pictures.
6. Watson A, Sasse A (1998) Measuring perceived quality of speech and video in multimedia conferencing applications. *Proc. ACM Multimedia*, Bristol, UK.
7. ITU-T Rec. P.920 (2000) Interactive test methods for audiovisual communications.
8. Sefiridis V, Ghanbari M, Pearson DE (1992) Forgiveness effect in subjective assessment of packet video. *Electronics Letters* 28(1): 2013-2014.
9. Aldridge R, Davidoff J, Ghanbari M, Hands D, Pearson D (1995) Measurement of scene-dependent quality variations in digitally coded television pictures. *IEE Proc. Vision, Signal and Image Processing* 142: 149-154.
10. ITU-T Rec. P.911 (1998) Subjective audiovisual quality assessment methods for multimedia applications.
11. Araujo P, Frøyland L (2004) Statistical approach to the rational selection of experimental subjects. *Accreditation and Quality Assurance* 10(5): 185-189.
12. Jumisko-Pyykkö S, Häkkinen J (2008) Profiles of the evaluators – Impact of psychographic variables on the consumer-oriented quality assessment of mobile television. *Proc. SPIE Multimedia on Mobile Devices*, vol. 6821, San Jose, CA.

13. Speranza F, Poulin F, Renaud R, Caron M, Dupras J (2010) Objective and subjective quality assessment with expert and non-expert viewers. *Proc. QoMEX*, Trondheim, Norway.
14. Köster O, Jessen M, Khairi F, Eckert H (2007) Auditory-perceptual identification of voice quality by expert and non-expert listeners. *Proc. International Congress of Phonetic Sciences*, Saarbrücken, Germany.
15. Choi H, Jeong T, Lee C (2009) Subjective video quality comparison using various displays. *Optical Engineering* 48(4): 037002.
16. Sullivan M, Pratt J, Kortum P (2008) Practical issues in subjective video quality evaluation: Human factors vs. psychophysical image quality evaluation. *Proc. uxTV*, Silicon Valley, CA.
17. Staelens N et al. (2010) Assessing quality of experience of IPTV and video on demand services in real-life environments. *IEEE Trans. Broadcasting* 56(4): 458-466.
18. ITU-T Rec. P.910 (2008) Subjective video quality assessment methods for multimedia applications.
19. ITU-T Rec. P.800 (1996) Methods for subjective determination of transmission quality.
20. ITU-R Rec. BS.1284 (2003) General methods for the subjective assessment of sound quality.
21. Corriveau P (2006) Video quality testing. In Wu R, Rao KR (eds) *Digital Video Image Quality and Perceptual Coding*, chapter 5, CRC Press.
22. Guilford JP (1954) *Psychometric Methods.* 2nd edition, New York: McGraw-Hill.
23. Tominaga T, Hayashi T, Okamoto J, Takahashi A (2010) Performance comparisons of subjective quality assessment methods for mobile video. *Proc. QoMEX*, Trondheim, Norway.
24. Rouse DM, Pepion R, Le Callet P, Hemami SS (2010) Tradeoffs in subjective testing methods for image and video quality assessment. *Proc. SPIE Human Vision and Electronic Imaging*, vol. 7527, San Jose, CA.
25. Pinson M, Wolf S (2003) Comparing subjective video quality testing methodologies. *Proc. SPIE VCIP*, vol. 5150, Lugano, Switzerland.
26. ITU-R Report BT.1082-1 (1990) Studies toward the unification of picture assessment methodology.
27. Mullin J, Smallwood L, Watson A, Wilson GM (2001) New techniques for assessing audio and video quality in real-time interactive communication. *Proc. IHM-HCI*, Lille, France.
28. Winkler S (2009) On the properties of subjective ratings in video quality experiments. *Proc. QoMEX*, San Diego, CA.
29. Huynh-Thu Q, Garcia MN, Speranza F, Corriveau P, Raake A (2011) Study of rating scales for subjective quality assessment of high-definition video. *IEEE Trans. Broadcasting* 57(1): 1-14.
30. Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81-97.
31. VQEG (2000) Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment.
32. Winkler S (2005) *Digital Video Quality: Vision Models and Metrics*. Wiley.
33. Xu J, Xing L, Perkis A, Jiang Y (2011) On the properties of mean opinion scores for quality of experience management. *Proc. IEEE International Symposium on Multimedia*, Dana Point, USA.
34. ITU-T Rec. G.107 (2011) The E-Model: A computational model for use in transmission planning.
35. ITU-T Rec. G.1070 (2007) Opinion model for video-telephony applications.
36. Brunnström K, Hands D, Speranza F, Webster A (2009) VQEG validation and ITU standardization of objective perceptual video quality. *IEEE Signal Processing Magazine,* 26(3): 96–101.
37. Lin W, Jay Kuo CC (2011) Perceptual visual quality metrics: A survey. *J. Visual Communication and Image Representation* 22(4): 297-312.
38. Chikkerur S, Sundaram V, Reisslein M, Karam LJ (2011) Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcasting* 57(2): 165-182.
39. You J, Reiter U, Hannuksela MM, Gabbouj M, Perkis A (2010) Perceptual-based quality assessment for audio–visual services: A survey. *Signal Processing: Image Communication* 25(7): 482-501.
40. ITU-T Rec. P.1401 (2012) Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.

41. ITU-T Rec. P.862.1 (2003) Mapping function for transforming P.862 raw result scores to MOS-LQO.
42. Mittal A, Muralidhar GS, Ghosh J, Bovik AC (2012) Blind image quality assessment without human training using latent quality factors. *IEEE Signal Processing Letters* 19(2): 75-78.
43. Xue W, Zhang L, Mou X (2013) Learning without human scores for blind image quality assessment. *Proc. CVPR*, Portland, OR.
44. Vranješ M, Rimac-Drlje S, Grgic K (2013) Review of objective video quality metrics and performance comparison using different databases. *Signal Processing: Image Communication* 28(1): 1-19.
45. Streijl R, Winkler S, Hands D (2010) Perceptual quality measurement – Towards a more efficient process for validating objective models. *IEEE Signal Processing Magazine* 27(4): 136-140.
46. Sen D (2011) Determining the dimensions of speech quality from PCA and MDS analysis of the Diagnostic Acceptability Measure. *Proc. MESAQIN*, Prague, Czech Republic.
47. Janowski L, Papir Z (2009) Modeling subjective tests of quality of experience with a Generalized Linear Model. *Proc. QoMEX*, San Diego, CA.
48. Carroll RJ, Wu CFJ, Ruppert D (1988) The effect of estimating weights in weighted least squares. *J. American Statistical Association* 83(404): 1045-1054.
49. Nachlieli H, Shaked D (2011) Measuring the quality of quality measures. *IEEE Trans. Image Processing* 20(1): 76-87.
50. Wu O, Hu W, Gao J (2011) Learning to predict the perceived visual quality of photos. *Proc. ICCV*, Barcelona, Spain.
51. Brooks AC, Zhao X, Pappas TN (2008) Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions. *IEEE Trans. Image Processing* 17: 1261-1273.
52. Wang Z, Simoncelli EP (2008) Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision* 8(12): 1-13.
53. Ciaramello FM, Reibman AR (2011) Systematic stress testing of image quality estimators. *Proc. ICIP*, Brussels, Belgium.
54. Reibman AR (2012) A strategy to jointly test image quality estimators subjectively. *Proc. ICIP*, Orlando, FL.
55. ATIS-0800041 (2010) Implementer’s guide to QoS metrics.
56. ATIS-0800008 (2011) QoS metrics for linear IPTV. Version 2.
57. ITU-T Rec. P.800.2 (2013) Mean Opinion Score (MOS) interpretation and reporting.
58. de Koning TCM, Veldhoven P, Knoche H, Kooij RE (2007) Of MOS and men: Bridging the gap between objective and subjective quality measurements in mobile TV. *Proc. SPIE Multimedia on Mobile Devices*, vol. 6507, San Jose, CA.
59. Jumisko-Pyykkö S, Malamal Vadakital VK, Hannuksela MM (2008) Acceptance threshold: A bidimensional research method for user-oriented quality evaluation studies. *Int’l J. Digital Multimedia Broadcasting* 712380.
60. Watson AB, Kreslake L (2001) Measurement of visual impairment scales for digital video. *Proc. SPIE Human Vision and Electronic Imaging*, vol. 4299, San Jose, CA.
61. Jia Y, Lin W, Kassim AA (2006) Estimating just-noticeable distortion for video. *IEEE Trans. Circuits and Systems for Video Technology* 16(7): 820-829.
62. Maloney LT, Yang JN (2003) Maximum likelihood difference scaling. *Journal of Vision* 3(8): 573-585.
63. Virtanen MT, Gleiss N, Goldstein M (1995) On the use of evaluative category scales in telecommunications. *Proc. Human Factors in Telecommunications*, pp. 253-260.
64. Preminger JE, Van Tassell DJ (1995) Quantifying the relationship between speech quality and speech intelligibility. *J. Speech and Hearing Research* 38: 714-725.
65. Voiers WD (1977) Diagnostic Acceptability Measure for speech communication systems. *Proc. ICASSP*, Hartford, CT.
66. Martens JB (2002) Multidimensional modeling of image quality. *Proceedings of the IEEE* 90(1): 133-153.
67. Borg I, Groenen P (2005) *Modern Multidimensional Scaling: Theory and Applications*. Springer.
68. Osgood CE, Suci GJ, Tannenbaum PH (1957). *The Measurement of Meaning*. University of Illinois Press.

69. Hinterleitner F, Norrenbrock CR, Möller S (2013) Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. *Proc. ISCA Speech Synthesis Workshop*, Barcelona, Spain.
70. Wältermann M (2013) *Dimension-based Quality Modeling of Transmitted Speech*. Springer.
71. Ghinea G, Thomas J (2005) Quality of perception: User quality of service in multimedia presentations. *IEEE Trans. Multimedia* 7(4): 786 - 789.
72. ANSI T1.801.02 (1996) Digital transport of video teleconferencing/ video telephony signals – performance terms, definitions, and examples.
73. Radun J et al. (2008) Content and quality: Interpretation-based estimation of image quality. *ACM Trans. Applied Perception* 4(4): 21.
74. Strohmeier D, Jumisko-Pyykkö S, Kunze K (2010) Open Profiling of Quality: A mixed method approach to understanding multimodal quality perception. *Advances in Multimedia* 2010: 658980.
75. Egger S, Ries M, Reichl P (2010) Quality-of-Experience beyond MOS: Experiences with a holistic user test methodology for interactive video services. *Proc. 21st ITC Specialist Seminar*, Miyazaki, Japan.
76. Yuen M, Wu HR (1998) A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing* 70(3): 247-278.
77. Hemami SS, Reibman AR (2010) No-reference image and video quality estimation: Applications and human-motivated design. *Signal Processing: Image Communication* 25(7): 469-481.
78. Reddy A, Estrin D, Govindan R (2000) Large-scale fault isolation. *IEEE J. Selected Areas in Communications* 18(5): 733-743.
79. Mahimkar A et al. (2009) Towards automated performance diagnosis in a large IPTV network. *Proc. ACM SIGCOMM*, Barcelona, Spain.
80. Keimel C, Rothbucher M, Shen H, Diepold K (2011) Video is a cube. *IEEE Signal Processing Magazine* 28(6): 41-49.
81. Venkataraman M, Chatterjee M (2009) Evaluating quality of experience for streaming video in real time. *Proc. GLOBECOM*, Honolulu, HI.
82. Zhai G, Cai J, Lin W, Yang X, Zhang W, Etoh M (2008) Cross-dimensional perceptual quality assessment for low bitrate videos. *IEEE Trans. Multimedia* 10(7): 1316-1324.
83. TM Forum GB917 (2012) *SLA Management Handbook*, Release 3.1.
84. Seo SS, Kwon A, Kang JM, Hong JWK (2011) OSLAM: towards ontology-based SLA management for IPTV services. *Proc. ManFI Workshop*, Dublin, Ireland.
85. Broadband Forum (2006) Triple-play service quality of experience (QoE) requirements. Technical Report TR-126.
86. ITU-T Rec. G.1080 (2008) Quality of experience requirements for IPTV services.
87. Cermak GW (2009) Consumer opinions about frequency of artifacts in digital video. *IEEE J. Selected Topics in Signal Processing* 3(2): 336-343.
88. Suresh N, Jayant H (2006) 'Mean Time Between Failures': A subjectively meaningful video quality metric. *Proc. ICASSP*, Toulouse, France.
89. Shi J, Zhou S (2009) Quality control and improvement for multistage systems: A survey. *IIE Transactions* 41(9): 744-753.
90. Knoche H, de Meer H, Kirsh D (1999) Utility curves: Mean Opinion Scores considered biased. *Proc. IWQoS*, London, UK.
91. McGurk H, Macdonald JW (1976) Hearing lips and seeing voices. *Nature* 264: 746-748.
92. Laghari KUR, Crespi N, Connelly K (2012) Toward total quality of experience: A QoE model in a communication ecosystem. *IEEE Communications Magazine* 50(4): 58-65.
93. Qualinet (2013) Qualinet white paper on definitions of quality of experience.
94. Winkler S (2012) Analysis of public image and video databases for quality assessment. *IEEE Journal on Selected Topics in Signal Processing* 6(6): 616-625.