International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.1401

(01/2020)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Statistical analysis, evaluation and reporting guidelines of quality measurements

Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models

Recommendation ITU-T P.1401

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
Methods for objective and subjective assessment of speech and video quality	P.800–P.899
Audiovisual quality in multimedia services	P.900–P.999
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	**P.1400–P.1499**
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.1401

Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models

Summary

A stable and self-sustained statistical evaluation procedure is required in the development of objective quality algorithms. This is required regardless of whether the algorithms will be used for estimating subscriber perception of voice, video, audio or multimedia quality. Recommendation ITU-T P.1401 presents a framework for the statistical evaluation of objective quality algorithms regardless of the assessed media type.

History

Edition	Recommendation	Approval	Study Group	Unique ID[*]
1.0	ITU-T P.1401	2012-07-14	12	11.1002/1000/11688
1.1	ITU-T P.1401 (2012) Cor. 1	2014-10-29	12	11.1002/1000/12327
2.0	ITU-T P.1401	2020-01-13	12	11.1002/1000/14159

Keywords

Audio, mean opinion score, MOS, multimedia, objective quality algorithms, QoE, quality models, statistical evaluation, subjective testing, video, voice.

[*] To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

Recommendation ITU-T P.1401

Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models

1 Scope

This Recommendation defines methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. This Recommendation can be used to assess any objective model that predicts a subjective judgement of a subjective test procedure. Guidance is provided on the design and cleansing of subjective test data, as well as the statistical metrics for model selection and characterization. Frameworks, metrics and example procedures are described. Specific procedures, minimum performance requirements or objectives to be used when selecting a model are not provided, as these depend on the scope of the model being assessed and are not part of this Recommendation. In this Recommendation, the term "sample" refers to any type of media, and the terms "model" and "algorithm" are interchangeable.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T G.107] Recommendation ITU-T G.107 (2015), *The E-model: a computational model for use in transmission planning*.

[ITU-T G.1070] Recommendation ITU-T G.1070 (2018), *Opinion model for video-telephony applications*.

[ITU-T J.247] Recommendation ITU-T J.247 (2008), *Objective perceptual multimedia video quality measurement in the presence of a full reference*.

[ITU-T J.341] Recommendation ITU-T J.341 (2016), *Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*.

[ITU-T P.563] Recommendation ITU-T P.563 (2004), *Single-ended method for objective speech quality assessment in narrow-band telephony applications*.

[ITU-T P.564] Recommendation ITU-T P.564 (2007), *Conformance testing for voice over IP transmission quality assessment models*.

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[ITU-T P.862] Recommendation ITU-T P.862 (2001), *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*.

[ITU-T P.862.1] Recommendation ITU-T P.862.1 (2003), *Mapping function for transforming P.862 raw result scores to MOS-LQO*.

[ITU-T P.863] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction*.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

[ITU-R BS.1116] Recommendation ITU-R BS.1116-3 (2015), *Methods for the subjective assessment of small impairments in audio systems.*

[ITU-R BT.500] Recommendation ITU-R BT.500-14 (2019), *Methodology for the subjective assessment of the quality of television images.*

3 Definitions

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CS	Circuit Switch
IMS	Internet protocol Multimedia Systems
IPTV	Internet Protocol TV
MOS	Mean Opinion Score
OR	Outlier ratio
rmse	root mean square error
VoIP	Voice over Internet Protocol

5 Conventions

None.

6 Subjective test and objective algorithms

Objective algorithms estimate subscriber perception and an algorithm's performance is evaluated against subjective test results. For this to be valid, the subjective tests must be well defined and accurate to avoid misinterpretations of algorithm accuracy.

6.1 Aspects related to subjective testing

For a stable evaluation to take place, a number of subjective tests are required. A group of subjective tests used in the evaluation of objective algorithms is called a subjective data pool. The subjective data pool must contain subjective tests that adhere to well established testing procedures, such as those listed below:

- Listening speech quality – See [ITU-T P.800]
- Video for multimedia – See [ITU-T P.910]
- Audiovisual for multimedia – See [ITU-T P.911]
- Quality of TV pictures – See [ITU-R BT.500]
- Audio and music – See [ITU-R BS.1116]

The rapid development of networks and their myriad services require that objective quality metrics take into account new technologies (such as codecs, bandwidth), new networks (such as long term evolution (LTE), IP multimedia subsystems (IMS)) and/or new applications (such mobile/IPTV, video streaming) and therefore must cope with new types of media degradations which impact the subscribers' perception of the quality. First, it is necessary to design subjective tests that can accurately capture the impact of these degradations on a subscriber's' perception. These subjective tests require performing comprehensive experiments that are consistent in their results. In the last several years, these types of tests have been developed and used for objective quality evaluation metrics needed to deal with new test conditions such as (re-)buffering in multimedia streaming, super-wideband voice and the evaluation of the combined audio-visual impact that is modelled by multimedia quality evaluation algorithms.

Subjective testing is extensively covered in [b-ITU-T Handbook]. For the purpose of this Recommendation only the main aspects of subjective testing are discussed. The following aspects are required for accurate evaluation of an objective quality algorithm:

i) Voters are recommended to be naïve subjects representing normal subscribers whose perception is estimated by the objective quality models. However, for specific applications such as new codec developments or voice enhancement device evaluations, experienced voters are more suitable.

ii) The number of voters per sample should meet the subjective testing requirements as described in the appropriate Recommendations, such as [ITU-T P.800], [ITU-T P.910], [ITU-T P.911], [ITU-R BT 500] or [ITU-R BS.1116]. Depending on the goal of the prediction (per sample prediction or per condition prediction) a minimum of 24 voters either per sample or per condition is recommended.

iii) The experiments performed, either in the same or different labs, could contain an anchor pool of samples that best represent the particular application under evaluation. This would ensure the experiment's alignment with respect to quality range and distortion types in the experiment, and would maintain consistency/repeatability across experiments and/or labs. However, it should be noted that even when anchor samples are used, a bias between different experiments is common. This is due to the fact that it is not always possible to include all distortion types in the anchor conditions.

6.2 Aspects related to objective algorithms

There are two main categories of objective algorithms. The first is based on network and device parameters (describing the network using abstract parameters). See [ITU-T P.564] (used as a voice quality evaluation tool), [ITU-T G.107] (used as a voice planning tool) and [ITU-T G.1070] (used as a video telephony planning tool). The second uses the real media signal (e.g., voice, video, audio) characteristics to describe the network performance, and the network is considered to be a black-box. Such models are often based on perceptual models. The analysis is not restricted to the media signal itself but can also take into account associated information, for example from the transport layer (often called 'hybrid models'). Both model categories estimate the media quality as subjectively perceived by test users. Regardless of their type, the evaluation procedure stays generally the same. However, the selection process of an algorithm depends on whether the standardization process is defined as a competition between several algorithms or a collaboration of multiple algorithms. In the first case, the evaluation is focused on the comparison of a set of algorithms while the latter case is focused on the comparison against a minimum performance pre-defined threshold.

There are several main aspects related to the objective quality algorithms that need to be considered for their consistent and stable evaluation:

i) An algorithm's accuracy is impacted by modelling imperfections. However, when evaluating the performance of an algorithm, the imperfections of the mean opinion score (MOS) panel come into play. Therefore, it is recommended that various statistical performance evaluation metrics are used, as described in clause 7.

ii) Each of the categories of algorithms mentioned above is expected to exhibit different accuracies. Parameter based algorithms are expected to be less accurate than methods having access to the media signal itself (voice, video, audio, multimedia). Thus, algorithms applying perceptual models to the signal are expected to be the most accurate.

iii) Evaluations need to be performed per application type as well as across all types of applications (e.g., mobile/IPTV, video streaming for video/audio/multimedia metrics; CS and/or VoIP, including VoIP over IMS, and narrow/wide/super-wideband for voice metrics) for which the algorithm(s) has been designed.

iv) Pre-defined minimum performance quality thresholds need to be established both per test application, as well as across applications. These thresholds are required whenever a single algorithm is developed rather than a competition between several algorithms. However, even in the latter case, it is recommended that these minimum requirements are used when defining a quality acceptance threshold.

v) For algorithms that separately estimate user perception of different types of degradations, evaluation of each degradation type, as well as an overall performance quality metric needs to be calculated. This type of complete performance analysis requires that each degradation type be associated with a subjective quality score.

7 Evaluation framework

7.1 Data preparation

An important component of the evaluation framework is the databases used for evaluating the algorithm's performance. Both the content and quality of the databases can drastically impact the evaluation results.

7.1.1 Test and validation databases

The main scope of the development of the objective quality algorithms is for network design/deployment, performance evaluation and/or monitoring. The objective metrics is meant to cover the two phases of the network's life; the design/deployment phase, which requires many simulations of different test conditions, and the evaluation/monitoring phase, which involves recording field measurements (i.e., recordings of real, live network scenarios).

It is recommended that the development (training, testing), validation and evaluation of the objective algorithms' databases evenly and equally cover both simulated and live network conditions. Generally simulations are used for the analysis of dedicated distortions in well-controlled situations and often occur during development of equipment. Live scenarios are more complex and cover a series of individual degradation types and interferences between them. These complex scenarios can hardly be simulated. For avoiding biases between simulations and real field data, the evaluation set should contain databases where simulated and real field data are mixed in the same experiment.

It is strongly recommended to select databases for evaluation that reflect the use cases of the objective prediction method. The database set selected should reflect by a sufficient proportion each use case. This could, for example, be a requirement to use a 50 per cent simulated vs. 50 per cent live-recorded databases for methods applicable in lab and field measurements as defined in [ITU-T P.863], but it could also be a requirement to use a larger amount of live data in case of monitoring systems, or in case of a focus on lab recordings or simulations only for methods in lab use. The same importance applies to even and equal distributions of targeted applications and test conditions within each experiment. The number of individual test conditions and applications should also reflect the expected occurrence used later in the method.

The quality of the test samples used in each subjective experiment should allow the voters to use the entire MOS scale. This will diminish the risk of bias towards one end or middle-range of the scale.

7.1.2 Data cleansing

It is recommended that prior to analysis, an inspection of the subjective and objective test data be performed to uncover any types of outliers.

One example is the time clipping effect, often seen in speech quality metrics. Severe time clipping conditions can generate a fairly large difference between the absolute category rating (ACR) listening-only subjective scores and the objective scores that rely on the comparison between the original and the degraded speech samples. It could be desirable to evaluate these types of situations within specially designed experiments, or to remove these outliers from the evaluation.

Another example is related to the video quality metrics where the ACR with hidden reference tests are used. Reference samples present in the test with a MOS rating less than 4 should be identified and the file examined. If it is determined that the poor MOS values for these reference samples are due to low quality, those files are recommended to be removed from the analysis.

In addition, conditions or databases showing subjective scores with unexpected confidence intervals, as compared with other conditions or databases obtained in experiments with similar designs, should be carefully analysed and a decision made on how these will be used in the evaluation process. Likewise, similar databases which show extremely different scores for the anchor samples should be carefully considered in the evaluation.

7.2 Analysis types

The evaluation procedure performed could be based on one of four types of analysis or on a combination of these. The selected analysis type is related to the application (e.g., voice, video, audio, multimedia) as well as the algorithm type (i.e., perceptual or parametric). The selection is recommended as follows:

i) Analysis per experiment is required for all types of algorithms and applications and involves comparisons between algorithms and/or against pre-defined minimum performance quality thresholds for each subjective experiment. Statistical evaluation metrics are calculated per experiment which equates with a standalone population.

ii) Analysis across experiments is required for all types of algorithms and applications and involves comparisons between algorithms and/or against pre-defined minimum performance quality thresholds across all subjective experiments. This analysis requires the calculation of aggregated statistical evaluation metrics that are valid across all databases. In other words, it is necessary to ensure that the applied statistics use correct statistical assumptions on the entire sample population (i.e., all samples from all experiments).

iii) Analysis per file can be algorithm type dependent, but is necessary when non-stationary live-recorded samples are used. In addition, this analysis is applied regardless of whether per experiment and/or across experiment evaluation is performed. In this case, each sample is characterized by a single objective score that estimates the average scores obtained from individual voter's scoring of the same sample.

iv) Analysis per condition can be algorithm type dependent, but it can only be used for simulated samples, where each live-recorded sample equates to a condition.

7.3 Prediction on a numerical quality scale

Prediction on a numerical quality scale plays an important role in the evaluation of an objective model's performance. A thorough understanding of the topic is required in order to accomplish an accurate performance evaluation of an objective model. This clause incorporates experience and findings accumulated over many years. More details can be found in [b-Berger].

7.3.1 Comparing MOS values of different experiments

Objective models are trained to predict quality as it can be obtained in subjective experiments. The subjective experiments are considered as references. However, the intra- and inter-experimental variations should be considered when developing and evaluating an objective quality predictor.

In a subjective experiment, such as a listening only test (LOT), a group of people scores a set of test samples. Usually, untrained 'naïve' listeners are invited to score, for example, voice samples in their native language, or video samples not of their cultural context.

The group of people remains the same during the experiment, and the stimuli, that is, the 'context', covers the same range of test conditions for everyone in the group. The average of the individual scores per-sample or per-test condition forms the MOS. The MOS is an average with a certain statistical uncertainty, since not all people will mark the sample with exactly the same score.

However, the MOS is usually considered as a single number describing the quality of the sample or test condition. Statistical uncertainty in terms of standard deviation or confidence intervals is provided as additional information. Although accidental, individual false-scores can occur, the relation of the MOS in this experiment can be considered 'true,' but only for this group and this test design. Individual scores also drive the confidence interval, which describes the uncertainty of the MOS values, but only in this experiment.

Normally, only the MOS values from the same test type can be compared. That is, the MOS values from a listening-only test cannot be compared to MOS values from a conversational test. Furthermore, the MOS values from a listening-test that use an ACR scale cannot be directly compared to the MOS values from a degradation category rating (DCR) experiment.

However, even when MOS values from the same test types are compared, some limitations are present. Each experiment is slightly different, even if the same participants were involved.

First, a score assigned by a listener or viewer is never always the same, even when an experiment is repeated with the same samples in the same representation order. This can be regarded as a type of noise that is overlaid on the scores.

Second, there is a short-term context dependency. Subjects are influenced by the short-term history of the samples they have previously scored. For example, following one or two poor samples, subjects tend to score a mediocre sample higher. If a mediocre sample follows very good samples, there is a tendency to score the mediocre sample lower. As a strategy to average out this short-term context dependency, the presentation order should be varied for the individual test participants. However, the statistical uncertainty remains.

Third, the largest difference between subjective experiments is caused by medium- and long-term context effects associated with the average quality, the distribution of quality and the occurrence of individual distortions. This is in addition to the set of conditions that is included in the experiment. For example, in an experiment that contains mainly low quality samples, people tend to score them higher, and vice versa. This is because people tend to use the entire set of scores offered in an experiment. Despite verbal category labels, people adapt the scale to the qualities presented in the experiment. In addition, individual distortions for samples presented less often are scored lower, as compared to experiments where samples are presented more often and people become more familiar with them. This can be seen as a mid-term context dependency, and reflects the effects caused by the design of the individual experiment.

Last, there are long-term dependencies. These reflect the general cultural behaviour of the individuals as to the exact interpretation of the category labels, the cultural attitude to quality, and to language dependencies. The daily experiences with telecommunication or media have the same importance. Here it is noted that quality experience, and therefore expectation, may change over time. As people become familiar with mobile codecs and their distortion, it becomes part of their daily experience.

The same telecommunication channels are almost noise-free today compared to the plain old telephone service (POTS) telephony of decades ago.

All of these effects lead to differences between individual experiments. These effects cannot be avoided but can be minimized by providing informative instructions, well-balanced test designs, a sufficient number of participants, and a mixed presentation order.

Such influences can lead to different MOS values for the same test condition even when one or more stimuli are identical in two different experiments. Biases in the scale interpretation can exist, for example, when participants in experiment A assign lower scores than those in experiment B due to the design of the experiment or the behaviour of the test group. Depending on the test context, different gradients in quality can arise due to the order of the stimuli that are presented or due to the different focuses of the experiments. These differences make it very complicated to directly compare MOS values from different experiments.

The diagrams in Figure 7-1 illustrate this problem. Within an ITU-T activity (focusing on ITU-T G.729 in 1995, available as [b-ITU-T P-Sup.23]), a set of 44 different test conditions was defined. Across these test conditions, voice samples in different languages were transmitted and scored in the individual laboratories by native subjects for each language. Although the test conditions, test design, and listening conditions were exactly the same, there were differences in the reported MOS values. The left graph in Figure 7-1 shows the results that were obtained in a Canadian laboratory with North American samples (Experiment 1) plotted on the x-axis, and the same experiment conducted in Japan with Japanese samples and listeners plotted on the y-axis (Experiment 2).

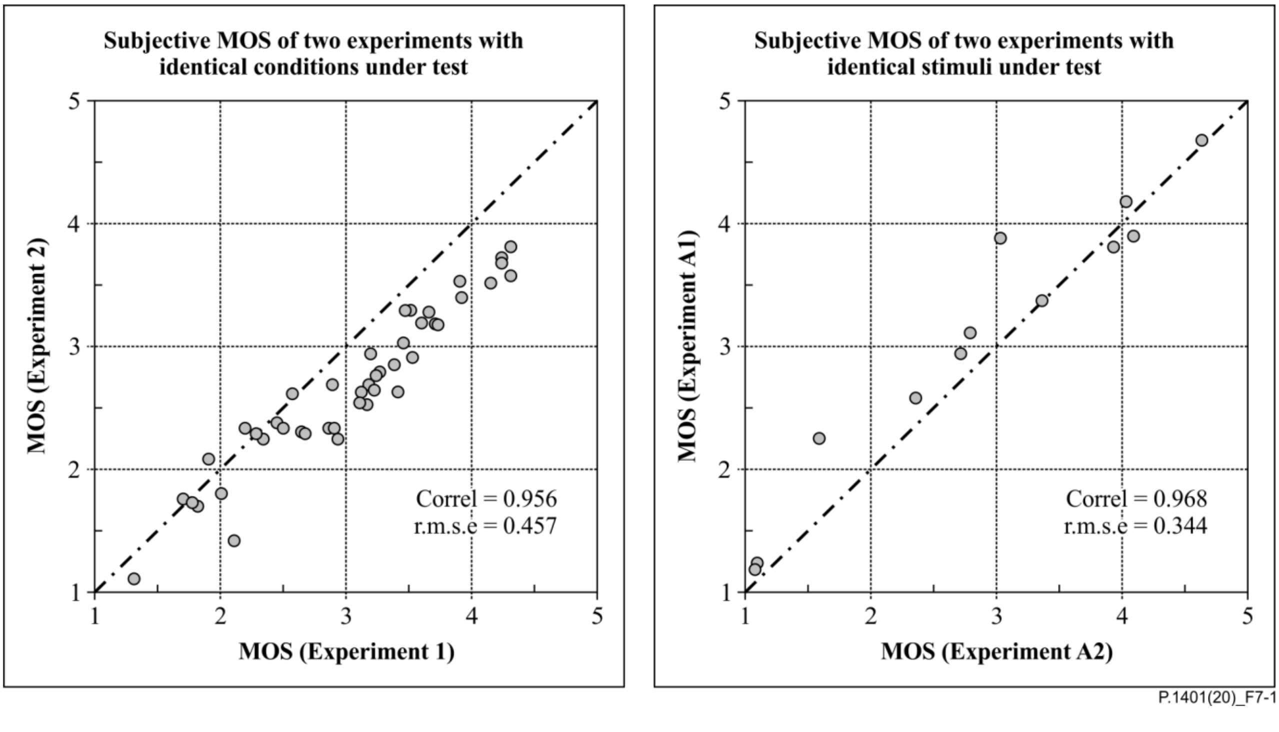


Figure 7-1 – Comparison of MOS values obtained in different subjective experiments

If the MOS values for each condition were the same, all points would be on the 45 degree line; however, the points are mostly below the line. This shows that in Experiment 2, a certain test condition was consistently assigned a lower score than in Experiment 1. However, it is not just a bias or a different gradient; there is also a kind of 'noise' and there are conditions where the qualitative relation is inverted.[1]

Such effects could easily be viewed as some type of language-dependence, however, this is not entirely the case. Considerable differences were also present when experiments were conducted in the same laboratory using identical stimuli, as seen in the results on the right graph of Figure 7-1. Both experiments were conducted in the same laboratory with the same test equipment and source samples, meaning that the texts and speakers, and the focus of the experiments were very similar. For formal reasons the group of the 24 test people were different.

Both experiments used identical voice stimuli as anchor conditions to provide an overlapping area to illustrate the differences. In principle, the qualitative ranking of the stimuli was similar in both experiments. However, there was a bias in the lower area and some larger differences in the MOS values even though the stimuli were identical.

Both examples illustrate how difficult it is to compare the MOS values of individual experiments even when the experiments follow the same guidelines. In addition to 'normal' uncertainties, systematically observed differences can be grouped into the following problem categories:

- **Bias or Offset**: A constant offset exists between the MOS values. This offset can be the result of the 'overall' quality that is presented during an experiment, which may influence participants to score more pessimistically or more optimistically. The offset can also be caused by different listening gear or environmental noises. A plain bias is quite rare to observe and is usually combined with a different gradient.
- **Different Gradient**: Relative quality distance between two identical stimuli or conditions is different during the experiments. In other words, the scores tend to become more pessimistic faster. This effect is usually caused by the test design, especially if the test does not have quality samples that cover the entire quality range. In such a case, testers tend to use the entire scale for the range of quality that is included.
- **Different Qualitative Rank Orders**: This problem category is the most severe. The main purpose of a subjective test is to determine the relative ranking of systems. For example, to show that the quality of A is better than B, which is better than C. This assumes that the relative quality ranking is constant and can always be reproduced.

  In practice, such a ranking cannot always be reproduced during other subjective tests.

  The MOS values always have a statistical uncertainty, which is usually expressed in the confidence interval. Serious analyses of subjective tests take this uncertainty into account and only rank the quality of A above the quality of B when the MOS values exhibit statistically significant differences. If the differences are not significant, the systems A and B are considered to be of equal quality. Since a confidence interval is usually in the range of 0.15 MOS, no finer resolution than 0.3 MOS in ranking can be achieved with a certain confidence.

[1] Regarding the two statistical values, the Pearson correlation coefficient appears quite high with 0.96, however, it implies a bias and gradient correction. The rmse shows the root mean square error of all differences as they can be seen in the diagram.

> Problems other than statistical uncertainties can lead to real changes in the relative quality ranking of systems. It is assumed that a subjective experiment will always determine the quality ranking of different systems correctly, even when the MOS values contain statistical uncertainties. The MOS values are always correct but they depend on the design of the experiment, for example, the distribution of distortion types present throughout the experiment. An under representation of a distortion leads to a relatively low MOS value while an over representation of a distortion leads to a high MOS value, as participants become desensitized to the distortion. Such context dependent effects may directly influence the quality ranking order. Other examples of context dependencies are unbalanced listening panels or a non-calibrated test setup.

A strategy to minimize scaling effects as biases and different gradients is to introduce defined anchor and reference conditions in two experiments, which can then be used to align the scores of the two experiments. In addition, design constraints are under discussion to make the distribution of distortion types and quality ranges comparable between different experiments in order to minimize rank order changes.

Ultimately, a subjective experiment remains a closed set. The experiment is self-contained and cannot be compared with other experiments.

7.3.2 Scale calibration of objective quality models

Objective models predict quality based on technical or physical information. Often partial results of individual analysis are combined into a single value in a late aggregation step. This single value is usually dimensionless and not tied to the numerical quality scale in subjective tests such as the one to five scale.

Since the prediction on such a scale is desired by users, transformation of the internal quality predictor value to a numerical quality scale is required. This scaling is usually the final step in an objective quality model. This scaling is an integral part of the objective model.

This scaling can involve multidimensional optimization against the statistical evaluation metrics across a large pool of media (e.g., voice, video, audio) samples carefully selected to uniformly cover all test conditions for which the algorithm has been designed.

The scaling procedure is created and proved by subjective reference experiments. Here the scaling is calculated such that the prediction widely follows the scale interpretation of the reference experiments, e.g., by choosing a scaling function that results in a minimum root mean square error (rmse) between the subjective reference experiments and the scaled objective predictions. The function and form of the applied scaling is defined by the range and form of the prediction values and to some extent by the chosen reference experiments. The scaling function can also be determined by technical requests such as desired upper limits.

A model, including proper scaling to a numerical scale, predicts this scale. Note that the interpretation of the scale is determined by the reference experiments chosen. That is, a test sample quality prediction is predicted in the same way as it is subjectively evaluated within the set of reference experiments.

As an example, an AMR 12.2 kbit/s sample may be scored in subjective experiment A at 4.2, in experiment B at 3.9 and in experiment C at 4.1. A model can be scaled to predict this sample at 4.07, the average across the reference experiments. However, the model does not predict the score of any individual experiment, only the generalized scale.

Therefore, the selection of reference experiments is essential to how the model uses or interprets the quality scale. Usually, a large number of well-balanced experiments are used to define the scaling of a model. It is also possible to design a smaller set of reference experiments to be used for scaling purposes.

In principle, after scaling, an objective model predicts a generalized numerical quality scale. Despite prediction inaccuracies, it predicts a certain test sample as it would be subjectively qualified on average across a wide number of subjective experiments.

In a competition, a training set is generally used for developing the model and this training set, or sub-set, is used during the development phase to define the scaling function for each model.

Later, in the evaluation phase, further experiments are conducted and made available. At the end of this stage, before a model becomes approved, it makes sense to recalculate the scaling function using the increased number of experiments available. This scaling adjustment does not affect the performance of the model, however it generalizes the scaling slightly more due to the increased number of reference experiments.

7.3.3 Performance evaluation of objective measures and compensation for the variance between subjective experiments

The work to develop, approve and recommend a final model can be divided into three phases (before/during/after approval):

1) Development and training phase; ends with submission of a model for evaluation.

2) Evaluation and selection phase; ends with approval of one or more models or with the rejection of model(s) in case of insufficient performance.

3) Characterization phase; characterizes the approved model in more detail.

In the second phase, statistical metrics are applied to characterize the performance of a model with respect to its prediction accuracy as related to a subjective experiment. The accuracy is usually determined by a statistical interpretation of the difference between the MOS values of the subjective test and its prediction by the model on a generalized scale. Usually, a set of subjective experiments is used and for each experiment statistical metrics based on these prediction differences.

Subjective experiments result in MOS values that are true and valid; a subjective test does not evaluate quality on a generalized scale like an objective predictor does. Despite inaccuracies of the objective model, there are always differences between the objective prediction and the subjective MOS values due to the scale differences. These differences are smaller the closer the experiment is to the average across all experiments used for the generalized scaling.

The calculated prediction difference is an overlay of the following:

1) The inaccuracy of the prediction;

2) The difference of the subjective MOS values, obtained in the individual experiment, and the average MOS values of the sample across many experiments.

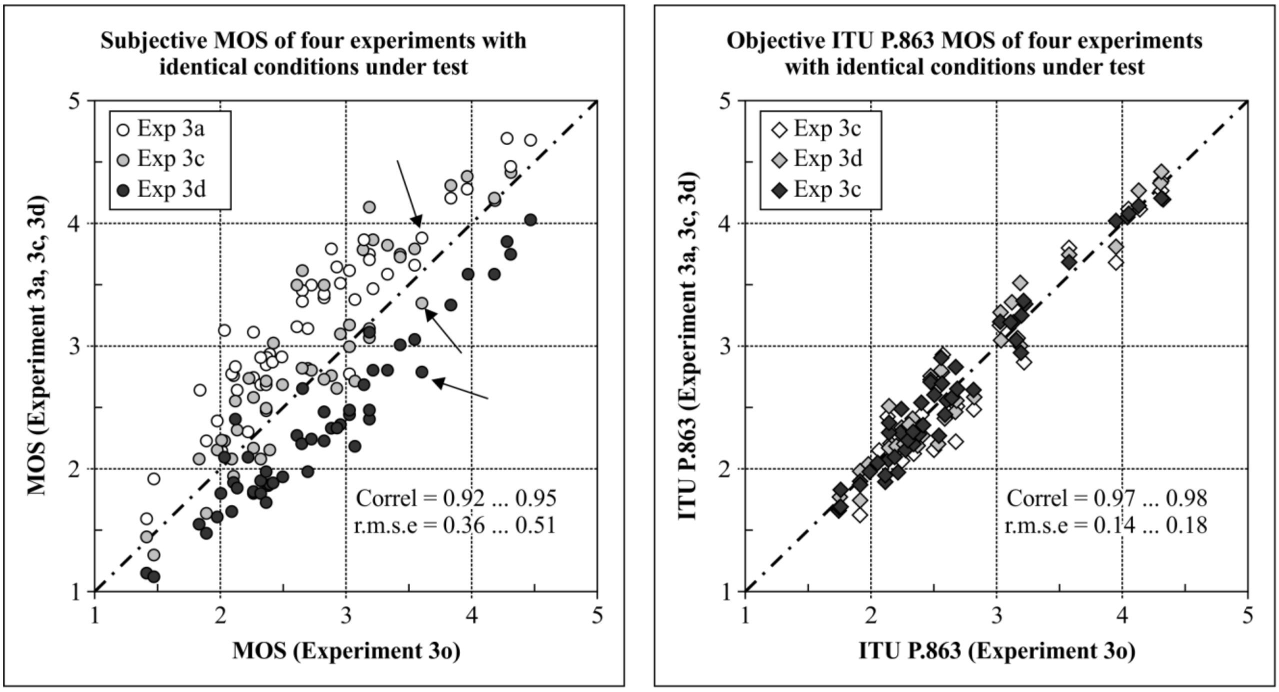
This results in some drawbacks. Even a perfect model will result in prediction differences due to the differences mentioned in 2) above. Model accuracy metrics are strongly dependent on the design of the experiments used for evaluation, and the further they are away from the reference set used for scaling, the larger the calculated prediction difference. The actual prediction inaccuracies are masked or compensated for by biased subjective experiments.

As a strategy to minimize this dependency on subjective experiments, an individual compensation is used. The basic assumption is that well-balanced and well-designed subjective experiments are reproducing the qualitative rank-order with high accuracy, while the actual scale range and the gradient, as explained above, may be subject to individual interpretation. Both can be compensated for by individual mappings, where bias and gradient become aligned towards a generalized scale as used by the objective model.

A MOS can differ between individual experiments for the training process of an objective model, even when same stimuli are used. This problem cannot be solved because an objective model cannot predict two or more different MOS values for the same signal. Instead, the training process that uses this large amount of data from different experiments leads to the prediction of a kind of average MOS for the experiments.

Figure 7-2 shows the test results for four experiments under exactly the same conditions (i.e., speech codecs, frame and bit errors, and background noise) that were carried out in four different languages and laboratories. These test results are available from [b-ITU-T P-Sup.23].

The diagram on the left in Figure 7-2 shows the MOS values of subjective experiments. The MOS values for American English are plotted on the x-axis while the values for French, Italian and Japanese are plotted on the y-axis. The latter three experiments are plotted and contrasted with the American English results.



P.1401(20)_F7-2

Figure 7-2 – Subjective and objective scores of experiments covering the same test conditions

The three arrows shown in the left diagram represent a single test condition used in all the experiments. The MOS value in the American English experiment (x-axis) is approximately 3.6, while the MOS values for the other experiments (y-axis) are at 2.8 (Exp. 3d), 3.4 (Exp. 3c), and 3.9 (Exp. 3a). Each experiment has a narrow distribution within itself; however, the scores of each experiment have different offsets. The large spread indicates considerable differences between the experiments.

The diagram on the right of Figure 7-2 shows results that were derived from the same experiments using objective quality predictions detailed in [ITU-T P.863] instead of MOS values. Note that these scores are much closer for each condition and do not have the same grouping per experiment as the subjective MOS values. The objective measure cannot predict the larger differences between the experiments that were caused by the different cultural attitudes, the interpretation of the scale labels, and the specific behaviour of the test group. The prediction is on a generalized scale. The processing conditions were identical, not the test samples, and thus small differences between the results remain.

As a consequence, an objective model that is trained using a large amount of data sets will never exactly match the MOS values of a single experiment. There will always be a difference between the MOS values and the objective model quality predictions.

Figure 7-3 shows the objective prediction scores plotted versus the MOS values for experiments 3c and 3d.

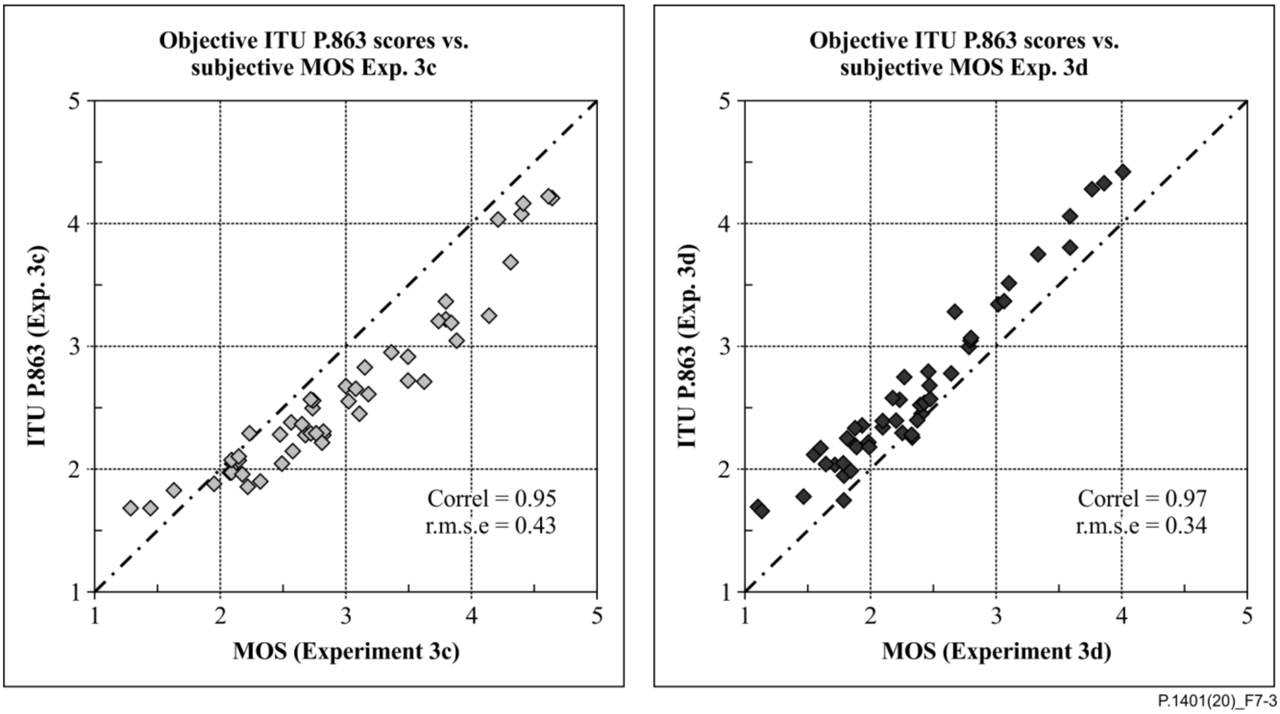


Figure 7-3 – Objective vs. subjective scores of two experiments of [ITU-T P-Sup.23]

The distribution is very narrow, which means that the reproduction of the rank-order is fairly accurate; however, there are some offsets from the 45 degree line. These offsets are not a problem or a malfunction of the objective model, but are caused by the inter-experiment differences of the subjective experiments. The objective model predicts a type of average across all of the experiments that are used in the training process. Consequently, the model cannot match individual scores when the distortion is the same.

The diagrams in Figures 7-2 and 7-3 use statistical measures to show the performance or accuracy of the objective measures. More specifically, Pearson's correlation coefficients and rmse values are plotted on the diagrams.

The difference between what appears to be true subjective MOS values and the outcome of the objective measure is often referred to as a prediction error. However, a more accurate term used to define this difference is 'prediction difference' without the methodological flaw of an always true MOS value.

This prediction difference is actually the difference between the predicted score and a MOS value taken from an individual experiment. Consequently, the prediction difference is influenced by the uncertainty of the objective measure and the differences between MOS values of the individual experiments.

An experiment-wise compensation of bias, without changing the qualitative rank-order, will minimize the problem of dependency on individual experimental setups.

Usually, a monotonous mapping function, a linear function or the more sophisticated monotonous part of a third order polynomial or a logistic function, is applied. The purpose of the mapping function is to minimize the rmse or another metric without changing the rank-order and by the optimal function of a given structure, that is often, first and third order polynomials. Mapping function examples can be found in [ITU-T P.862], [ITU-T P.862.1], [ITU-T P.563], [ITU-T P.863], and ITU-T J.341]. The mapping function compensates for offsets, different biases, and other shifts between the scores, without changing the rank order. The function is usually applied to the predicted scores before the statistical metrics are calculated.

Figure 7-4 illustrates the effect of mapping on the data set from experiment 3c.

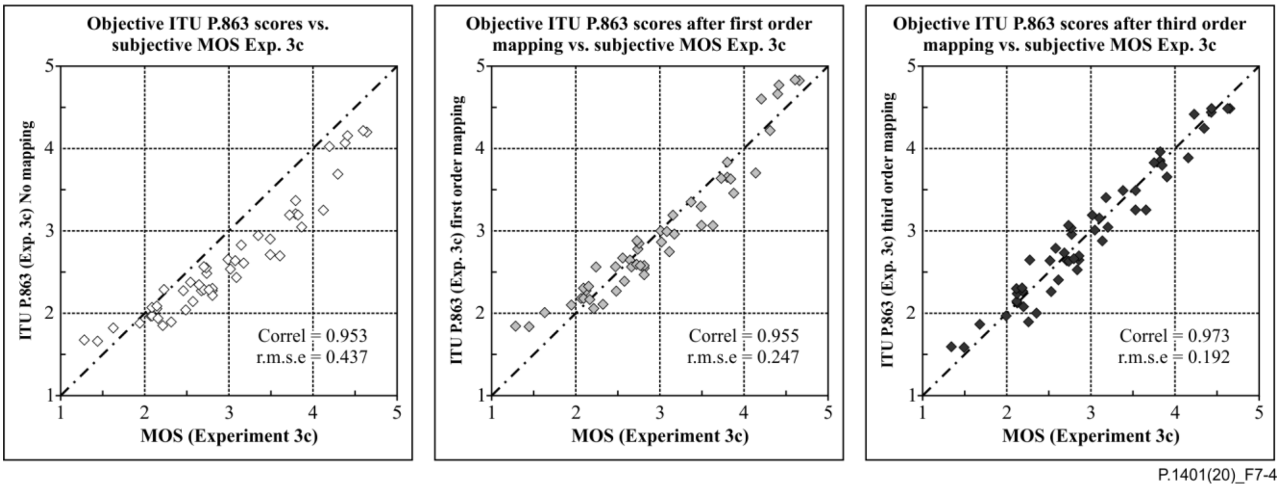


Figure 7-4 – Objective vs. subjective raw scores, after first-order and after third-order mapping

The diagram on the left shows the relation between the objective raw scores, that is, [ITU-T P.863], and the subjective MOS values. There is a clear offset and a slightly different gradient. The data appear to form a banana shape.

The scatterplot in the middle applies a linear mapping function to the objective scores. The equation has the form:

$$y' = a + by, \text{ where } rmse(x, y') \rightarrow min$$

As a result, the offset is compensated and the gradient is closer to the 45 degree line. Consequently, the rmse drops from 0.44 down to 0.25. The Pearson correlation coefficient remains the same since the coefficient uses an implicit first-order mapping.

The scatterplot on the right shows the results after applying a third order polynomial function. The equation has the form:

$$y'' = a + by + cy^2 + dy^3, \text{ where } rmse(x, y'') \rightarrow min \text{ and } f(y) = \text{monotonous between } y''_{min} \text{ and } y''_{max}$$

This mapping corrects the offset and the gradient, and linearizes the banana shape. Due to the improved linearization, the rmse drops to 0.19 for this experiment while the Pearson correlation coefficient increases to 0.97. The rank order of the scores remains the same. The part of the third order polynomial that was used was chosen by monotonous constraints.

The "mapping" discussed here is recommended in order to compensate for the possible variance between several subjective experiments. The comparison of different models is required for the performance evaluation process. Well designed and balanced experiments which use common anchor samples are expected to show very small differences. However, they are not completely avoidable.

In general, the alteration is not statistically significant. If the selected mapping function does not appropriately describe the dependency between the objective and subjective scores, then the performance evaluation process is flawed and its results are misleading. Appendix I presents a demonstration case of the mapping effect on an algorithm's performance.

When a third-order mapping is selected, a first-order mapping function can be used to evaluate any possible (artificial) gain that the cubic function may contribute to the algorithm's performance. Appendix II describes how this gain can be calculated for one of the statistical evaluation metrics, respectively the epsilon insensitive rmse described in clause 7.5 below. Similar procedures can be used for any of the metrics described below.

When using a mapping to the subjective scores, then the mapping must be applied per experiment, and before performing the calculation of the statistical evaluation metrics.

The application of a mapping applied before calculation of performance criteria must be stated to interpret the results accordingly. When applying no mapping function, the statistical performance values are biased by the difference of the individual experiments used, as compared to the average of the reference set in the evaluation of the model.

7.4 Uncertainty of subjective results

An objective model is not expected to predict an average subjective opinion more accurately than an average test subject. Therefore, the statistical performance evaluation of objective quality evaluation algorithms could consider the uncertainty of the subjective votes by taking into account the MOS panel confidence intervals. Usually, the standard deviation and its corresponding confidence interval (ci95) are calculated, in order to determine the degree of uncertainty of the subjects' votes. These statistical parameters are based on a "file-based" or "condition-based" analysis to determine the uncertainty of the subjects per file, or per test condition. A test condition consists of a set of files processed under the same technical circumstances (i.e., applying the same distortions).

The reasons for the average subject's uncertainty of opinions are various and application dependent (i.e., audio/speech, video, multimedia). In case of speech for example, a specific talker or sentence that the talker says might offend some in a listener group while others might be pleased by the voice and the content of the file. Listeners may be biased by the different talkers and the file content, and a system under test might also have its 'preferences'. An example is when low pitched voices are better encoded than high pitched voices. The subjects might vote accordingly, but within a test condition a large variability of the average opinion on that condition may be observed. Consequently, in a file-based analysis the two effects above can be easily identified and separated. Contrarily, in a condition-based analysis the two effects will be mixed-up, thus making it difficult for an objective model to predict a solid opinion.

Similarly, in different video cases, (e.g., video content, video resolution, bit rate), the same phenomenon is expected.

Appendix III describes how subjective confidence intervals can be calculated in various test scenarios.

These confidence intervals are recommended to be taken into account when using them in the calculation of the statistical evaluation metrics, especially for per-files evaluation when the number of votes is generally below 30 and for MOS values positioned at the scale's ends. Random distributions with more than 30 samples can be approximated by normal distributions for which Gaussian assumptions can be statistically and safely applied. If fewer than 30 samples are used, then the normal distribution starts to become distorted and calculation of confidence intervals based on normality assumptions are no more valid. It can be shown how especially at the ends of the MOS scale, the distribution of individual scores per file becomes skewed. In the cases with fewer than 30 samples, the t-Student distribution can be used to calculate the confidence intervals as discussed in clause 7.7.

In addition, especially in the case of few samples per individual MOS values (generally occurring in a per-file analysis), a normality test could be considered in order to understand the data skewness against the normal distribution. The normality test is presented in Appendix IV.

7.5 Statistical evaluation metrics

The recommended statistical metrics for objective quality assessment need to cover three main aspects: accuracy, consistency and linearity against subjective data.

It is recommended that the prediction error be used for accuracy, the outlier ratio (OR) or the residual error distribution for consistency and the Pearson correlation coefficient for linearity. Detailed descriptions of these metrics are presented below. In addition, this clause provides the formulas for these metrics' confidence intervals as well as the statistical significance tests required for the comparison of these metrics calculated for different algorithms to be compared.

It should be noted that these metrics are calculated per experiment (test database). In addition, they do not take into account the subjective uncertainty. Therefore, the use of controlled and uniform test databases (e.g., generally using same number of votes per-files, respectively per condition) is recommended.

7.5.1 Absolute prediction error (rmse)

The accuracy of the objective metric is evaluated using the rmse evaluation metric.

The difference between measured and predicted MOS is defined as the absolute prediction error (Perror), as shown in Equation 7-1:

$$Perror(i) = MOS(i) - MOS_p(i) \tag{7-1}$$

where the index *i* denotes the speech sample.

The rmse of Perror is calculated with the formula shown in Equation 7-2:

$$rmse = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} Perror(i)^2} \tag{7-2}$$

where N denotes the number of samples; while the division to (N–1) ensures an unbiased estimator for the rmse.

The rmse is approximately characterized by a χ^2 (n), where n represents the degrees of freedom and is defined by Equation 7-3:

$$n = N - d \tag{7-3}$$

where N represents the number of samples and d = 4 denotes the degrees of freedom of the mapping function (3rd order polynomial function).

Using the χ^2 (n) distribution, the 95 per cent confidence interval for the rmse is given by Equation 7-4:

$$\frac{rmse \times \sqrt{N-d}}{\sqrt{\chi^4_{0.975}(N-d)}} < rmse < \frac{rmse \times \sqrt{N-d}}{\sqrt{\chi^4_{0.025}(N-d)}} \tag{7-4}$$

7.5.2 Residual error distribution and outlier ratio

As mentioned above, either the residual error distribution or the outlier ratio can be used to evaluate the consistency of the models.

7.5.2.1 Residual error distribution

The residual error distribution (as defined in Equation 7-1) was used previously for the speech quality metrics evaluation [ITU-T P.862]. The residual error's distribution exhibits some limitations. First, it is difficult to associate with its 95 per cent confidence intervals, which need to be calculated per-bin.

Second, it is difficult to perform a statistically significant comparison between models for this metric due to an expected lack of data for the upper bins. Therefore, the use the cumulative distribution (CDF) of the residual error is recommended. The probability of exhibiting residual errors below a pre-established threshold, Equation 7-5, is easily determined based on the CDF.

$$P(Perror < Perror, th) = Pth = \frac{Nth}{N} \tag{7-5}$$

Nth denotes all samples for which the residual error (Equation 7-1), remains below the imposed threshold and N represents the total number of samples used for the analysis.

Therefore, the probability Pth represents the proportion of samples exhibiting errors below the threshold in the total number of samples N. Thus, the binomial distribution can be used to characterize the probability Pth. The probability Pth of exhibiting errors below the threshold is represented by a distribution of proportions [b-Spiegel] characterized by the mean, Equation 7-6, and standard deviation, Equation 7-7:

$$P = Pth = \frac{Nth}{N} \tag{7-6}$$

$$\sigma_{Pth} = \sqrt{\frac{Pth \times (1 - Pth)}{N}} \tag{7-7}$$

In addition, statistical significance tests can be straightforwardly applied for the probability Pth, as is shown later on.

7.5.2.2 Outlier ratio

As mentioned, an algorithm's consistency can be evaluated using the outlier ratio (OR) which represents the number of "outlier-points" to total points N.

$$OR = \frac{TotalNoOutliers}{N} \tag{7-8}$$

where an outlier is defined as a point for which the error (Equation 7-1) exceeds the 95 per cent confidence interval of the mean MOS value, as described in Equation 7-9.

$$\left|Perror(i)\right| > \frac{z \times \sigma(MOS(i))}{\sqrt{Nsubj}} \tag{7-9}$$

where σ(MOS(i)) represents the standard deviation of the individual scores associated with the media sample i, and Nsubj is the number of voters per media sample i. The 95 per cent confidence interval limit defined by the variable z is determined based on Nsubj. If Nsubj>30, then the Gaussian distribution can be used, and therefore z=1.96. If Nsubj<30, the t-Student distribution is used and the variable z becomes variable t and its value depends on the Nsubj, respectively the degrees of freedom df=Nsubj–1.

Note that the threshold defining an outlier can be changed depending on the desired strength for the accuracy's requirements; thresholds higher than the 95 per cent confidence interval allow more lose requirements.

The OR represents the proportion of outliers in N number of samples. Thus, the binomial distribution can be used to characterize the outlier ratio. The OR is represented by a distribution of proportions [b-Spiegel] characterized by the mean, Equation 7-10, and standard deviation, Equation 7-11:

$$p = or = \frac{TotalNoOut\ liers}{N} \tag{7-10}$$

$$\sigma_{or} = \sqrt{\frac{or \times (1 - or)}{N}} \tag{7-11}$$

For N>30, the binomial distribution, which characterizes the proportion p, can be approximated with the Gaussian distribution. Therefore, the 95 per cent confidence interval of the OR is given by Equation 7-12:

$$\pm 1.96 \times \sigma_{or} \tag{7-12}$$

NOTE – If fewer than N<30 samples are used, then the Gaussian distribution is replaced by the t-Student distribution with the variable t depending on the number of samples, respectively the degrees of freedom (df) [b-Spiegel].

Note that due to the fact that the OR depends on the standard of subjective scores as well as on the number of voters used in the subjective tests, the ratio is sensitive to the quality scale of the tested samples and the entire design of the subjective test. It is expected that experiments, which are not very well balanced may be misleading for comparing results from different experiments. An experiment which contains quality scores mainly in the upper and/or lower end of the quality scale are expected to exhibit lower MOS standard values. More balanced experiments could show higher MOS standard values. Therefore, an OR that emerged from the comparison of a balanced and a less balanced experiment may be difficult to achieve.

Due to the outlier's sensitivity, the use of this metric in the characterization phase, rather than during the validation and evaluation processes of the objective algorithm, is recommended.

7.5.3 Pearson correlation coefficient

The Pearson correlation coefficient R, Equation 7-13, measures the linear relationship between a model's performance and the subjective data.

$$R = \frac{\sum_{i=1}^{N} (Xi - \overline{X}) \times (Yi - \overline{Y})}{\sqrt{\sum (Xi - \overline{X})^2} \times \sqrt{\sum (Yi - \overline{Y})^2}} \tag{7-13}$$

Xi denotes the subjective score MOS and Yi the objective one (MOSp). N represents the total number of speech samples considered in the analysis.

It is known [b-Spiegel] that the statistic z, Equation 7-14, (also called Fisher z transformation) is approximately normally distributed and its standard deviation, Equation 7-15, are defined by:

$$z = 0.5 \cdot \ln\left(\frac{1+R}{1-R}\right) \tag{7-14}$$

$$\sigma_z = \sqrt{\frac{1}{N-3}} \tag{7-15}$$

The 95 per cent confidence interval for the correlation coefficient is determined using the Gaussian distribution, which characterizes the variable z and is given by Equation 7-16:

$$Z \pm 1.96 \times \sigma_z \tag{7-16}$$

NOTE – If fewer than N<30 samples are used, then the Gaussian distribution needs to be replaced by the two-tailed t-Student distribution with t depending on the degree of freedom [b-Spiegel].

7.6 Statistical significance evaluation

The statistical evaluation requires the comparison of the metrics mentioned in clause 7.5 for different objective models. This comparison needs to be based on the statistical significance tests described in clauses 7.6.1 to 7.6.5. As already mentioned in clause 7.3.2, if more than two algorithms are compared, a multiple-comparison method needs to be used, as described in clause 7.6.5.

7.6.1 Significance of the difference between the correlation coefficients

This test is based on the assumption that the normal distribution is a good fit for the speech quality scores' populations. The statistical significance test for the difference between the correlation coefficients uses the H_0 hypothesis that assumes that there is no significant difference between correlation coefficients. The H_1 hypothesis considers that the difference is significant, although not specifying whether the difference is better or worse.

The test uses the Fisher-z transformation, Equation 7-14 [b-Spiegel]. The normally distributed statistic, defined in Equation 7-17, is determined for each comparison and evaluated against the 95 per cent t-Student value for the two-tail test dependent on the number of degrees of freedom df.

$$Z_N = \frac{z1 - z2 - \mu_{(z1-z2)}}{\sigma_{(z1-z2)}} \qquad (7\text{-}17)$$

where

$$\mu_{(z1-z2)} = 0 \qquad (7\text{-}18)$$

and

$$\sigma_{(z1-z2)} = \sqrt{\sigma_{z1}^2 + \sigma_{z2}^2} \qquad (7\text{-}19)$$

σ_{z1} and σ_{z2} represent the standard deviation of the Fisher-z statistic for each of the compared correlation coefficients. The mean, defined by Equation 7-18, is set to zero due to the H_0 hypothesis and the standard deviation of the difference metric z1-z2 is defined by Equation 7-19. The standard deviation of the Fisher-z statistic is given by Equation 7-20 [b-Spiegel]:

$$\sigma_z = \sqrt{1/(N-3)} \qquad (7\text{-}20)$$

where N represents the total number of samples used for the calculation of each of the two correlation coefficients.

If the achieved ZN statistics is below the two-tailed t value then hypothesis H0 is true. If ZN is larger than this value, then the hypothesis H1 is true.

7.6.2 Significance of the difference between the outlier ratios

The OR, defined by Equation 7-8, is described by a binomial distribution of parameters (p, 1-p), where p is defined by Equation 7-10. In this case P is equivalent with the probability of success of the binomial distribution.

The distribution of differences of proportions from two binomially distributed populations with parameters (p1, 1-p1) and (p2, 1-p2) (where p1 and p2 correspond to the two compared ORs) is approximated by a normal distribution for N1, N2 >30, with the mean:

$$\mu_{(p1-p2)} = \mu(p1) - \mu(p2) = p1 - p2 = 0 \qquad (7\text{-}21)$$

and standard deviation:

$$\sigma_{p1-p2} = \sqrt{\frac{\sigma(p1)^2}{N1} + \frac{\sigma(p2)^2}{N2}} \qquad (7\text{-}22)$$

The null hypothesis in this case considers that there is no difference between the population parameters p1 and p2, respectively p1=p2. Therefore, the mean, Equation 7-21, is zero and the standard distribution, Equation 7-22, becomes Equation 7-23:

$$\sigma_{p1-p2} = \sqrt{p \times (1-p) \times (\frac{1}{N1} + \frac{1}{N2})} \quad (7\text{-}23)$$

where N1 and N2 represent the total number of samples of the compared ORs p1 versus p2. The variable p is defined by Equation 7-24:

$$p = \frac{N1 \times p1 + N2 \times p2}{N1 + N2} \quad (7\text{-}24)$$

Similar to the hypothesis test for the correlation coefficients, the normalized statistics ZN is calculated as:

$$Z_N = \frac{p1 - p2 - \mu_{(p1-p2)}}{\sigma_{(p1-p2)}} \quad (7\text{-}25)$$

and compared to the tabulated t value of t-Student distribution for the 95 per cent significance level of the two tailed test. If the calculated ZN > t, then the compared ORs p1 and p2 are statistically significant different, with 95 per cent significance level.

7.6.3 Statistical significant difference between probabilities of exhibiting errors below a pre-defined threshold

The statistical significance in this case is evaluated similar to the procedure applied for the OR. In this case the OR is replaced by the probability threshold Pth as defined in clause 7.5.2.2.

7.6.4 Significance of the difference between the root mean square errors

Assuming that the two populations are normally distributed, the comparison procedure is similar to the one used for the correlation coefficients. The H_0 hypothesis considers that there is no difference between rmse values. The alternative H_1 hypothesis assumes that the lower prediction error value is statistically significantly lower. The statistics defined by Equation 7-26 have an F-distribution with n1 and n2 degrees of freedom [b-Spiegel].

$$q = \frac{rmse^2{}_{\max}}{rmse^2{}_{\min}} \quad (7\text{-}26)$$

The rmse,max is the highest rmse and rmse,min is the lowest rmse involved in the comparison. The q statistic is evaluated against the tabulated value F(0.05, n1, n2) that ensures a 95 per cent significance level. If the calculated value for q is larger than the tabulated value, then the difference between the compared rmse is significant. The n1 and n2 degrees of freedom are given by N1-n (n=4 for a third order polynomial mapping, n=1 for a first order polynomial mapping), respectively and N2-n, with N1 and N2 representing the total number of samples for the compared average rmse (prediction errors).

7.6.5 Significance test in the case of multiple comparisons

When comparing more than two objective quality algorithms, the number of statistical tests that are performed should be taken into consideration. This is required in order to control the risk of committing Type-I errors, that is, to incorrectly conclude that a significant difference exists, while it has just appeared by chance. The risk increases with the number of comparisons that are performed [b-Brunnström].

The Bonferroni method [b-Maxwell] may be used to control Type-I errors, where the considered significance level (α) must be divided by the number of comparisons (n) so that the significance level for each comparison will be α⁄n. For example, if there are 10 comparisons and the overall α = 0.05, then each comparison must have a significance level of 0.05⁄10 = 0.005.

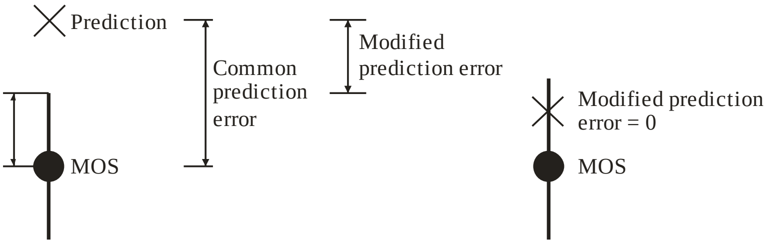
There are different scenarios for which the number of comparisons can grow very rapidly. If *X*, different algorithms should be compared with each other, then all pairwise comparisons will be *X(X-1)/2*. For example, for 10 algorithms, this will be 45 comparisons, and for 20 algorithms, it will be 190 comparisons. The corresponding α for 95% confidence will then be α = 0.0011 and α = 0.00026 for each comparison, respectively.

For extreme cases when the number of comparisons becomes very large (> 1000) the Bonferroni method can be too conservative, and more efficient methods may be needed, such as Holm [b-Holm], which is as safe as Bonferroni, but more efficient. For a very large number of comparisons (e.g., data mining), other methods that control the false discovery rate may be better to use, see e.g., [b-Benjamini-1] and [b-Benjamini-2].

7.7 Statistical evaluation in the context of subjective uncertainty: epsilon insensitive rmse and its statistical significance

As mentioned above, the statistical metric for accuracy's evaluation is based on the absolute prediction error Perror (Equation 7-1).

When the uncertainty of the subjective scores is taken into account, a statistical metric called ***epsilon-insensitive rmse*** (***rmse****) can be used**.** This is a rmse that considers the confidence interval of the individual MOS scores. This metric is calculated like the traditional rmse, but small differences to the target value are not counted. This rmse considers only differences related to an epsilon-wide band around the target value. This 'epsilon' is defined as the 95 per cent confidence interval of the subjective MOS value. By definition, the uncertainty of the MOS is taken into account in this evaluation. The rmse* is calculated on a prediction error as illustrated in Figure 7-5 below.



P.1401(12)_F7-05

Figure 7-5 – Calculation of rmse*

This modified rmse can be described as follows:

$$Perror(i) = \max(0, |MOSLQS(i) - MOSLQO(i)| - ci_{95}(i)) \qquad (7\text{-}27)$$

where the index *i* denotes the condition or the sample.

It should be noted that for MOS values situated at the ends of the scale, and in the case in which fewer than 30 votes have been used to calculate an individual mean MOS value, use the t-Student distribution for determining the CI% confidence intervals:

$$95\%CI = \pm t \times \sigma(MOSi) \qquad (7\text{-}28)$$

where i=1...N denotes the number of samples used for determining the individual mean MOS value and the variable t of the t-Student distribution is determined based on the number of degree of freedom, df=N-1.

The final modified *rmse** is calculated as usual but based on *Perror* with the formula:

$$rmse^* = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N} Perror(i)^2} \qquad (7\text{-}29)$$

where the index *i* denotes the condition or the sample, *N* denotes the number of conditions or samples and *d* the number of freedoms. The degree of freedom *d* is set to 4 in case a 3rd order mapping is applied in prior to the values, *d* is set to 2 in case of a 1st order mapping.

The rmse* is calculated per database and it gives an impression of how the prediction error exceeds the Ci95.

Since the rmse* is a modified form of the traditional rmse, some statistical assumptions that are valid for rmse may not apply. This might be the case if on average the difference between measured and predicted MOS is smaller or equal to the size of the confidence interval, see equation (7-27). Under the assumption of normal distribution, the statistical significance of the difference between the two rmse* values is calculated like it is in the traditional case. Assuming that the two populations are normally distributed, the comparison procedure is like the one used for the correlation coefficients. The H_0 hypothesis considers that there is no difference between the rmse* values. The alternative H_1 hypothesis assumes that the lower prediction error value is statistically significantly lower. The statistics q defined by the equation below has a F-distribution with n1 and n2 degrees of freedom [b-Spiegel].

$$q = \frac{(rmse*)^2{}_{\max}}{(rmse*)^2{}_{\min}} \qquad (7\text{-}30)$$

rmse*,max is the highest rmse* and rmse*,min is the lowest rmse* involved in the comparison. The q statistic is evaluated against the tabulated value F(0.05, n1, n2) that ensures a 95 per cent significance level. The n1 and n2 degrees of freedom are given by N1–n (n=3 for a third order polynomial mapping; n=2 for a first order polynomial mapping), respectively and N2–n. N1 and N2 represent the total number of samples for the compared average rmse* (prediction errors).

7.8 Statistical evaluation of the overall performance

The overall performance of a model is defined by its performance across each experiment (i.e., test database) as well as across all experiments. Therefore, results per experiment should be aggregated in an overall figure of merit. In this paragraph a figure of merit is recommended.

7.8.1 Databases weighting

Depending on the type of database (i.e., speech, video, audio or audio-visual) and its considered importance, the databases used in the evaluation process may have different weights. For example, in the case of speech, newly created databases focused on new technologies (e.g., VoIP over IMS in the voice case) may have higher weights than databases containing older technologies. With multimedia cases, audio-only and video-only databases may be considered with different weights depending on the aimed application.

Therefore, each experiment contributes different weights in the calculation of the aggregated statistical evaluation metric representing the figure of merit of each algorithm.

7.8.2 Aggregated statistical significant distance measure

Using a score represented by a defined distance is a well-known method in the graphs theory when searching for the minimum resistance path. The Viterbi coder (Rake receiver) is an example of applying this method. The same idea is used in this case.

Based on the statistical significance definition of each of the metrics (see clause 7.5), the algorithms can be compared using a statistical significance distance measure (SSDM) calculated for each model on a per experiment basis. The SSDM represents the figure of merit of a model per experiment and can be calculated as follows:

$$d_{k,v} = \sum_{i=1}^{Nmetric} W(i) * \max(0, StatMetricF(0.05, N_k, N_k)Result$$

where *StatMetricF(0.05,Nk,Nk)Result* denotes the result of the statistical significance test for each evaluated metric i=1...Nmetric (e.g., correlation coefficient, OR, rmse). The index k denotes the experiment, while index *v* denotes the objective model. F(0.05, n1, n2) is the tabulated value of the F-distribution for n1 and n2 degrees of freedom and 95 per cent significance level. *Nk* describes the number of considered samples (files or conditions) in experiment *k*. The function W(i) represents the weight that is allocated based the importance of each metric to the evaluation. In addition, the sum of the weights should round up to 1. If all metrics are considered equally important, then all weights equal 1. The W(i) = 0, when statistical metric i is not considered in the evaluation for a particular reason. The importance of one statistic versus another is based on whether the algorithms have been optimized towards that particular metric(s).

The result of the statistical significance test is defined by:

$$StatMetricF(0.05, N_k, N_k) = Z_N - F(0.05, N_k, N_k)) \tag{7-31}$$

where ZN represents the normalized Fisher statistics obtained for each metric's statistical significance test case.

The overall performance for an algorithm *v* is defined as follows:

$$p_v = \sum_{k=1}^{M} w_k \times d_{k,v} \tag{7-32}$$

where *M* is the total number of databases across the sets, *k* is the index of the database, $d_{k,v}$ is distance measure of the model *v* for the database *k*, and where w_k represents the weight of the database *k* as calculated by:

$$w_k = \frac{W_s}{N_s} \tag{7-33}$$

where *s* defines the set from which the database *k* is from, W_s is the weight for the set *s* (as defined above) and N_s is the number of database in the set *s*.

7.8.3 Statistical significance of the aggregated SSDM

To determine if two or more models are statistically equivalent or not, a statistical significance test is applied to the aggregated distance measures p_v available for all models.

The value p_v is the aggregated distance for *v* model and p_{min} is lowest p_v in the evaluation. The value *K* describes the degree of freedom that is N-4 for a third order polynomial mapping and N-2 for a first order mapping; N represents the total number of databases covered by all sets.

$$t_v = \max\left(0, \frac{p_v}{P_{min}+c} - F(0.05, K, K)\right) \tag{7-34}$$

If $t_v = 0$, the model *v* is considered as statistically equivalent to the model with $p = p_{min}$. If $t_v > 0$, the model *v* is considered as significantly statistically worse than the lowest $p = p_{min}$. Due to proved calculations for the speech quality assessment metrics' case, the constant c is recommended to be set to 0.0004.

8 Guidance on algorithm selection

Selecting a best-performing media quality evaluation algorithm can be complex and challenging and depends on a variety of factors. Some of these are:

- media type, (e.g., speech/audio/video/audio-video),
- model type (e.g., parametric, perceptual with all its flavours: full/reduced/non-reference, hybrid),
- scope (e.g., one single output representing the estimation of the perceived overall quality; set of outputs describing the estimation of a set of perceived media degradations),

– continuous influence of new technologies,

– subjective testing design keeping pace with the new technologies,

– approach used for the model development (e.g., competition, collaboration),

– parties involved.

It is out of scope of this Recommendation to cover all these scenarios. Rather, this Recommendation provides guidance that should be considered when selecting the best performing model/algorithm.

8.1 Per experiment performance

The performance per experiment is reflected by the SSDM metric. The algorithms with statistically-equal lowest SSDM values per experiment perform the best for that particular experiment. It should be noted that depending on the context of the evaluation (e.g., media type, algorithm type), not all statistical metrics are required to be considered for the evaluation.

8.2 Overall figure of merit

The overall performance is defined by the overall figure of merit calculated as the aggregated ssdm across all experiments as described above. The best performing algorithms should exhibit the lowest statistically-equal figure of merits.

8.3 Worst performance cases

The evaluation process should take into account the worst-performing case for all evaluated algorithms. In order to perform this evaluation, consideration of all four types of analysis: per experiment, across experiments, per file, and per test are required.

In addition the worst-performing case should regard all the statistical metrics considered in the evaluation process. It is recommended that a careful analysis be made and a decision taken on an algorithm which is the best performer, but shows as the worst case in at least one instance (e.g., per one experiment).

8.4 Averaging statistical metrics across experiments

Although not recommended, averaging statistical metrics across experiments per model can be used to evaluate overall performance. However, statistical limitations of the averaging process do not provide a clear statistical meaning. However, it is mentioned in the context of this Recommendation since it has been used several times for media quality evaluation algorithms. See [ITU-T P.862], [ITU-T P.863] and [ITU-T J.247].

Appendix V presents such a case for the epsilon insensitive rmse, rmse* [ITU-T P.863].

9 Special cases

9.1 Evaluation of algorithms with more than one output

The objective quality evaluation metrics designed to estimate the subscriber's perception of various dimensions of quality degradations (e.g., blurriness and blockiness in video; or loudness and coloration in voice), are required to be performed for each degradation type, as well as on the overall performance. The evaluation framework described above and the statistical metrics should be applied in this special case.

9.2 Evaluation of algorithms against pre-defined minimum performance requirements

As mentioned in clause 6.2, the selection process of an algorithm depends on whether the standardization process is defined as a competition between several algorithms or a collaboration of multiple algorithms into a single model. In the latter case, the evaluation framework is as described above, but the comparison evaluates a single algorithm's performance against a pre-defined minimum performance threshold.

Therefore, the minimum performance thresholds must be defined. It is recommended that these be defined based on previous experiences, whenever available. These scenarios include the case of either a new or improved standard, or a parametric (including planning) or hybrid model when a perceptual model is already in place.

In this special case, the evaluation framework described above and the statistical metrics should be used. The main difference is that the role of the "best performing model" is played by the minimum performance thresholds defined *a priori* to the evaluation process.

9.3 Scenarios using prediction error as unique statistical metric and datasets have different number of samples

For the sake of simplicity, it is common to use just one statistical metric for algorithm comparison. Prediction error is the metric generally used in this case since this is the one which best describes the performance of an algorithm against a subjective panel.

In addition, it is also common that training/testing and validation data sets are considered with different weights. This is done since it important to test the model on unknown data rather than training data, and hence apply more weight to the prediction error on the validation dataset. Also, individual datasets (subjective tests) can have different number of samples, which should be accounted for when computing an overall prediction error.

In such scenarios, the procedures outlined in clauses 9.3.1 and 9.3.2 must be used.

9.3.1 Aggregated performance measure

The evaluation of the algorithms is based on their performance across all experiments, included in the training (known) and validation (unknown) datasets. Therefore, an aggregated distance measure is calculated across all databases, taking into account individual weighting factors (e.g., W_training=0.1, w_validation=0.9). The performance of a model is measured using the aggregated error measure of equation (9-1), measuring the aggregated prediction error. Therefore, a better performance is equivalent to a smaller aggregated error.

The aggregated error measure *p* for the model/output *v* is defined as follows:

$$p_v = \frac{1}{W}\sum_{k=1}^{M} w_k \times RMSE_{k,v}^2 \tag{9-1}$$

where M is the total number of databases across the sets, k is the index of the database, $RMSE_{k,v}$ is the root mean square error for the model *v* for the database *k* and where w_k represents the weight of the database *k*. The normalisation constant W is given by $= \sum_{k=1}^{M} w_k$.

9.3.2 Statistical significance of the aggregated performance measure

The statistical significance test should be applied to the aggregated performance measures p_v of equation (9-1) available for all models as basis for consideration of two or more models as statistically equivalent or not.

The aggregated performance p_v is approximately chi-squared distributed according to the Welch-Satterthwaite approximation [b-Benjamini-2], with the degrees of freedom ϑ calculated by

$$\vartheta \approx \frac{\left(\sum_{k=1}^{M} w_k\right)^2}{\sum_{k=1}^{M} \frac{(w_k)^2}{\vartheta_k}} \tag{9-2}$$

where w_k represents the weight of the database k and ϑ_k denotes the degrees of freedom of $RMSE^2_{k,v}$ and is given by $\vartheta_k = N_k - 2$ and, with N_k the number of samples in the database k.

The value p_v is the aggregated performance for model v and p_{min} is the lowest p_v value in the evaluation. Then, the statistical significance test takes the form

$$t_v = \max(0, \frac{p_v}{p_{min}} - \mathrm{F}(0.95, \vartheta, \vartheta)) \quad (9\text{-}3)$$

If $t_v = 0$ the model v is considered as statistically equivalent to the model with $p = p_{min}$. In case that $t_v > 0$ the model v is considered as statistically significantly worse than the lowest $p = p_{min}$.

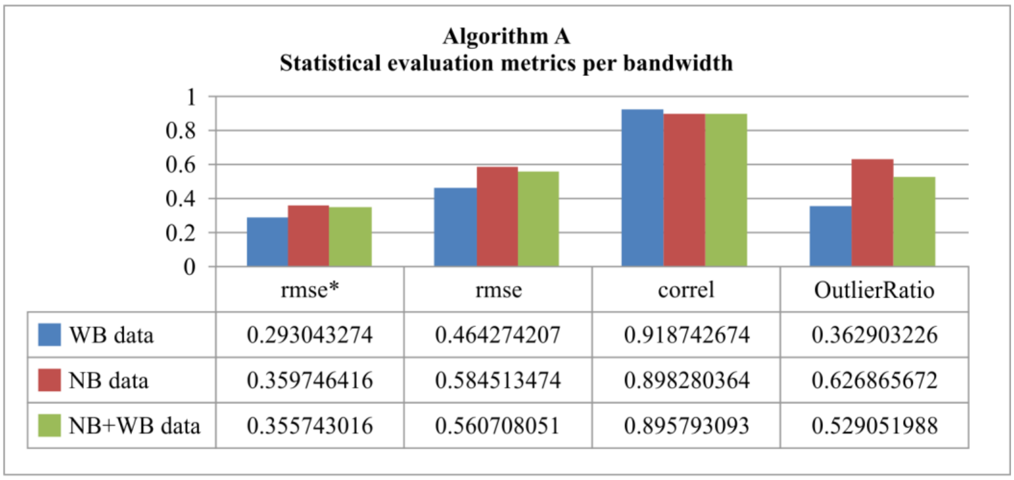
10 Demonstration cases

A demonstration case, using some of the metrics described above, is presented in this clause. Aspects related to their usage and interpretations are also discussed. Note that this demonstration case focuses on the aspects related to the evaluation metrics themselves, rather than the algorithm A performance on the test scenarios. Algorithm A is an objective metric for speech quality evaluation.

The performance behaviour of algorithm A on a set of scenarios is evaluated based on four statistical metrics: rmse, rmse*, Pearson correlation coefficient, and OR. The description of the data and the results are presented in Table 10-1. Figure 10-1 shows the results of the same metrics and the same data which have been grouped from another view of the test scenarios.

Table 10-1 – Statistical performance metrics

Scenario	rmse*	rmse	PearsonCorrCoef (%)	OutlierRatio (%) @ 95% CI-MOS panel
Overall	0.323	0.525	90.985	21.858
VoIPoverLTE-NBLive	0.454	0.691	89.479	38.298
CDMANB-Live (CS)	0.244	0.468	71.706	20.000
WCDMA-AMRNB-Live (CS)	0.165	0.355	70.900	6.897
WCDMA-AMRWB-Live (CS)	0.173	0.361	92.059	5.000
VoIPwithTS-NBLive	0.319	0.558	95.135	23.684
Sim TS-WB	0.590	0.795	71.050	37.500
Wireless – Live (CS)	0.200	0.402	84.439	36.697



	rmse*	rmse	correl	OutlierRatio
🟦 WB data	0.293043274	0.464274207	0.918742674	0.362903226
🟥 NB data	0.359746416	0.584513474	0.898280364	0.626865672
🟩 NB+WB data	0.355743016	0.560708051	0.895793093	0.529051988

P.1401(20)_F10-1

Figure 10-1 – Statistical performance metrics

The following conclusions are derived:

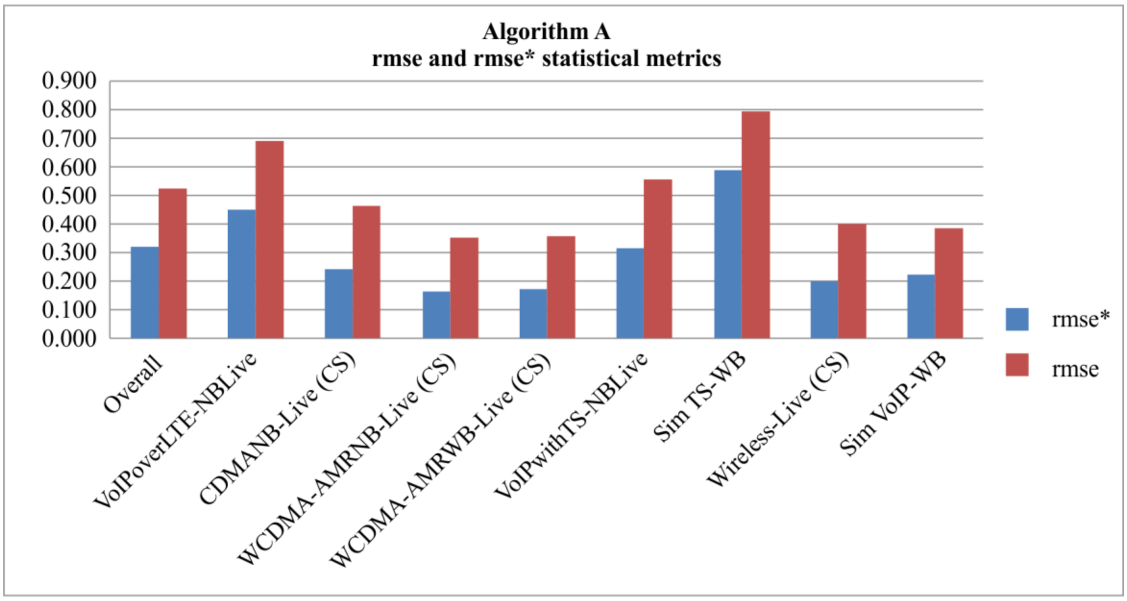
i) As expected, all metrics reflect the same performance behaviour for algorithm A.

ii) The rmse values show statistically poorer algorithm A performance than the rmse* metric, across all tested scenarios (Figure 10-1). This is, to a certain extent expected since the MOS panel 95 per cent confidence intervals were quite tight (i.e., roughly less than 0.3MOS). In other words, the rmse* which takes into consideration the subjective uncertainty, can smooth out the algorithm's imperfections in cases for which the MOS panel uncertainty is low (i.e., specific to a single subjective experiment and greater than 30 voters per file).

iii) The Pearson correlation coefficient (Figure 10-2) reflecting the algorithm's linearity with the subjective scores can be an important metric taken into consideration for an in-depth performance understanding. As seen in Table 10-1, there are scenarios (e.g., CDMA NB) for which the rmse value stays within a good performance range, while the correlation coefficient could drop below certain pre-defined accepted thresholds.

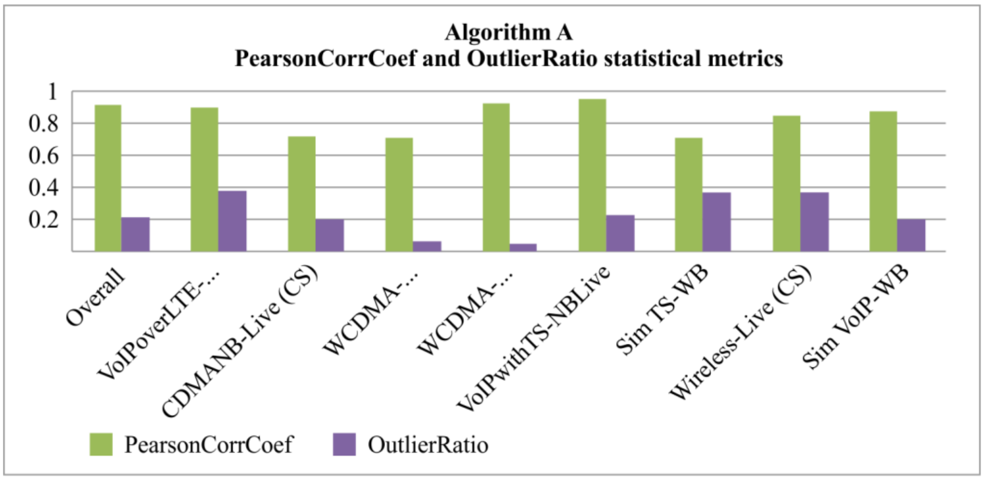
iv) The OR (Figure 10-3) is an important metric especially when needed to evaluate if the algorithm performs "good enough" for certain types of application. Challenges emerge from two sources. One is the definition of "good enough" and the other is the threshold for the OR definition. Both could drastically change the outcome of the analysis. Therefore, having a good understanding of the scope of the analysis is recommended when this metric is used.

v) For this particular demo case, if the OR=30 per cent then the applications VoIP over LTE-NB Live, Sim TS, and wireless CS –Live (Table 10-1 and Figure 10-3) would be of concern, and detailed analysis at per-individual score level may be required.



P.1401(20)_F10-2

Figure 10-2 – Rmse and rmse* values for various databases



P.1401(20)_F10-3

Figure 10-3 – Pearson correlation coefficient and outlier ratio for various databases

Appendix I

Algorithm mapping to the subjective scale

(This appendix does not form an integral part of this Recommendation.)

A score-mapping process is required to remove subjective test biases when comparing the performance of an objective model across many subjective tests and with other models. A mapping does not change an algorithm's rank-order accuracy but is required to make comparing objectively predicted scores with subjective scores meaningful. For more information on why mappings are required refer to clause 7.3 in this Recommendation.

A simplified example is presented first to illustrate how mappings influence algorithm selection. This is followed by the presentation of algorithm performance metrics for a single, real experiment.

In the simplified example, it is assumed that just one source of subjective test bias exists, i.e., that introduced by running an experiment twice. Table I.1 below shows the expected results.

Table I.1 – Simplified example expected results

	Experiment A MOS	Experiment B MOS
Condition A	4.0	4.3
Condition B	3.0	3.1
Condition C	2.0	1.9
Condition D	1.0	0.75

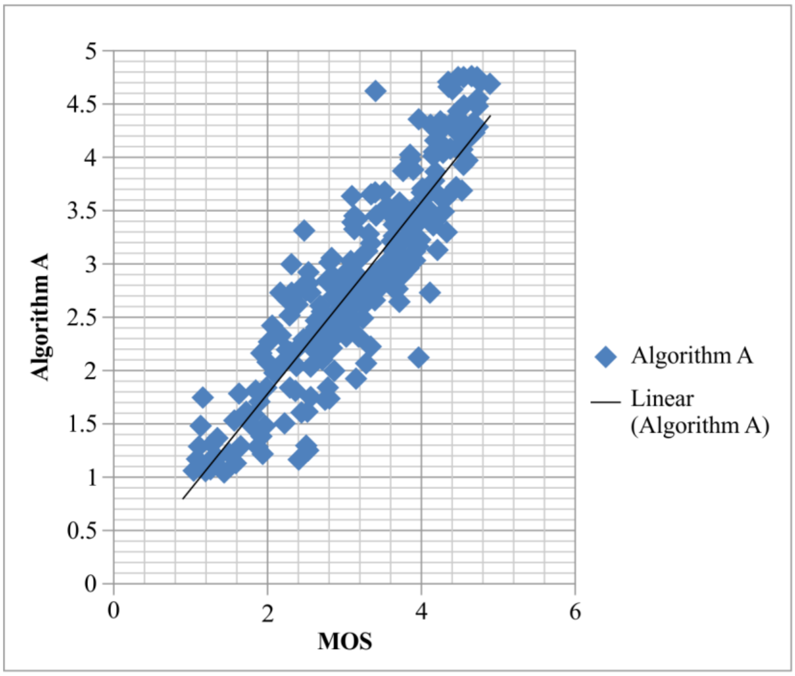
The scores from Experiment A do not match exactly with the scores from Experiment B. It is assumed that a number of algorithms (U to Z) are available to predict different values for each condition. Table I.2 below presents the predicted scores, and an rmse figure for each algorithm.

Table I.2 – Simplified example predicted scores and rmse per algorithm

	Algorithm U	Algorithm X	Algorithm Y	Algorithm Z
Condition A	4	4.3	4.15	3.9
Condition B	3	3.1	3.05	2.8
Condition C	2	1.9	1.95	1.7
Condition D	1	0.75	0.87	0.62
rmse	0.13	0.13	0.09	0.27

From this table, where no mappings are applied, it appears that algorithm Y performs the best, and algorithm Z performs the worst. However, in this example algorithm Z has predicted the same scores as algorithm Y, offset by 0.25 MOS. If this offset is removed using a simple linear mapping, then algorithms Y and Z perform equally. Clause 7.3.2 explains that before a model is approved it is typical to recalculate its final mapping, and thus if algorithm Z had a systematic offset over all experiments, it would be removed at this stage.

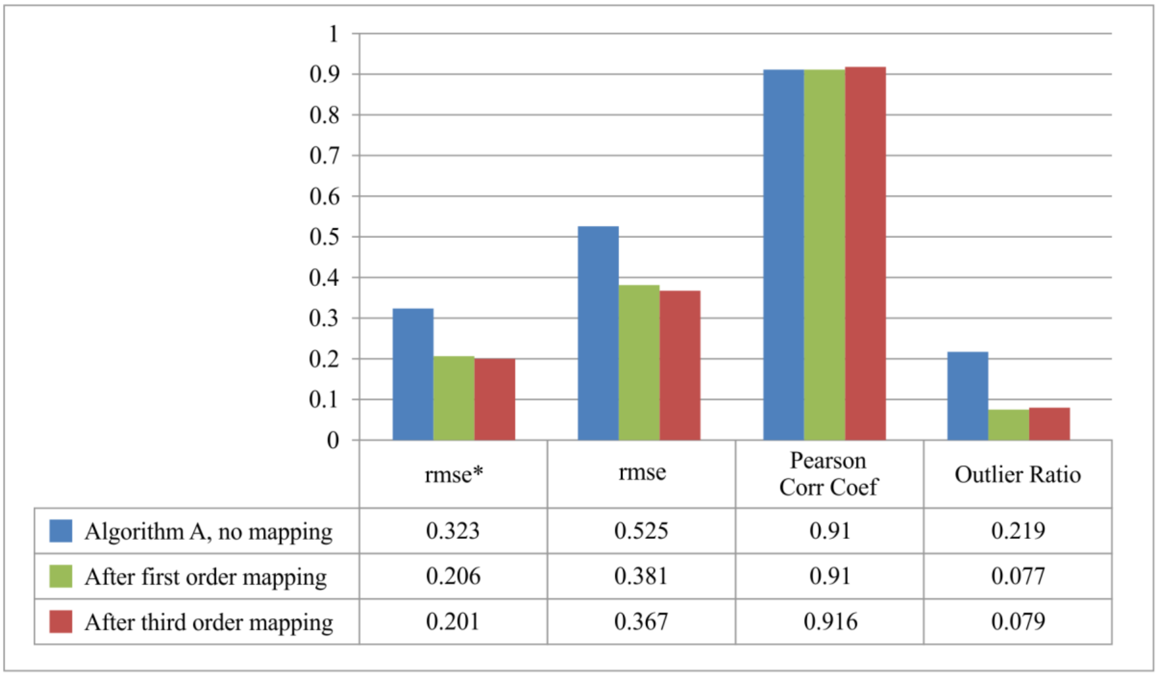
Although this example is contrived, this type of effect is often seen with real datasets. Figure I.1 shows how predictions from the real algorithm A match with a subjective experiment containing 350 samples.



P.1401(20)_FI.1

Figure I.1 – Objective to subjective scores dependency

Figure I.2 presents the performance-analysis metrics with no mapping, a first order linear mapping and a third order mapping. As expected, the Pearson correlation coefficient remains largely unaffected by the mapping, while the rmse and rmse* are significantly reduced.



	rmse*	rmse	Pearson Corr Coef	Outlier Ratio
🟦 Algorithm A, no mapping	0.323	0.525	0.91	0.219
🟩 After first order mapping	0.206	0.381	0.91	0.077
🟥 After third order mapping	0.201	0.367	0.916	0.079

P.1401(20)_FI.2

Figure I.2 – Statistical performance metrics

It is worth noting that in this example there is very little reduction in rmse and rmse* between the first and third order mappings. This reveals that the bias between the subjective test scores and the algorithm's predictions is largely a gradient and offset shift.

Appendix II

The impact of the third order versus first order mapping

(This appendix does not form an integral part of this Recommendation.)

II.1 Application of third order and first order mappings

The relationship between model outcomes and subjective MOS scores is usually slightly biased or slightly non-linear. This is often caused by the individual subjective test setups. It is a usual practice to compensate for these offsets and shapes by applying a third order mapping function to the raw outcomes of a model prior to further evaluation.

On the other hand, a third order mapping may mask systematic mis-prediction of a model and can lead to an overly optimistic model performance evaluation in these cases.

For this reason, the following statistics will be applied to model outcomes after 3rd order mapping and just after first order mapping in two separate threads.

Each objective model uses its own coefficients for these mapping functions. For the model and database the coefficients a0, a1, a2 and a3 for ONE third order mapping and the coefficients b0 and b1 for ONE first order mapping can be used. These mapping functions are applied to the raw outcomes of the models. Any condition averaging, or other aggregations of the scores, are applied to the mapped values. Thus, the same mapping functions are used for the primary and secondary analysis.

The third order mapping function must be monotonic within the entire interval of the scores it is applied to. It must be monotonic from the smallest to the highest predicted score for that database.

There may be some problems with mapping functions with very shallow or very steep slopes. A revision of the mapping functions (first and third order) can be applied after selection. Problematic conditions/use cases of the affected model(s) can be reported in the standard.

II.2 Gain of third order mapping

The non-linearity of the third order mapping function is proven by a gain derived in comparison to the scores following first order mapping. For each model and experiment an *rmse*_gain* is calculated that describes the gain in *rmse**.

For this calculation two sets of objective scores will used:

- MOSLQO after non-linear mapping
- MOSLQO after linear mapping

Based on each set of MOSLQO, the $rmse*_{k,v}$ is calculated for both sets of MOSLQO. That is, the resulting $rmse*^{3rd}_{k,v}$ as well as a $rmse*^{1st}_{k,v}$ are calculated. The derived gain is computed as follows:

$$rmse*\_gain_v = \sqrt{\left(\frac{1}{N}\sum_N \max\left(0, rmse*^{1st}_{i,v} - rmse*^{3rd}_{i,v}\right)^2\right)}$$

where *i* describes the experiment and *N* denotes the number of experiments. The index *v* determines the individual model.

To determine if two or more models are statistically equivalent or not, a statistical significance test is applied to the *rmse*_gain*$_v$ available for all considered models where *rmse*_gain*$_{min}$ is the lowest *rmse*_gain*$_v$ in that evaluation. *T* describes the degree of freedom that is the total number samples (files or conditions) in all considered experiments. The constant *c* is set to 0.01.

$$r_v = \max(0, \frac{rmse*\_gain_v^2}{\left(rmse*\_gain_{\min}^2 + c\right)} - \mathrm{F}(0.05, T\ , T\ ))$$

Appendix III

Confidence intervals calculation

(This appendix does not form an integral part of this Recommendation.)

This appendix describes the use-case of how MOS panel uncertainty can be taken into account and how the MOS confidence intervals of individual media (e.g., speech, audio, video) samples can be calculated in different test scenarios. This use-case is very well suited for situations where the test databases are, to a large extent, heterogeneous (e.g., different votes per sample, different number of samples per condition, etc.).

The confidence interval of subjective scores

The 95 per cent confidence interval (ci95) is usually calculated for all individual scores for a single file or test condition. The standard deviation σ and the number of individual scores M determines the confidence interval. Use of the accurate t-value for the given M is recommended.

$$ci_{95} = t(0.05, M)\frac{\sigma}{\sqrt{M}} \qquad \text{(III-1)}$$

Depending on whether a file-based analysis or a condition-based analysis is performed, the calculation of the standard deviation σ is computed as described below. The number of scores *M* is then replaced by the number of votes per-file *K* in the file-based analysis. In the condition-based analysis *M* is replaced by the number of votes per condition *N*.

III.1 The standard deviation for file-based analysis

The standard deviation σ_j of the individual votes $v_{j,l,k}$ of listener *k* for an audio file spoken by talker *l* and processed by condition *j* is defined as follows:

$$\sigma_{j,l} = \sqrt{\frac{\sum_{k=1}^{K}\left(v_{j,l,k} - MOSLQS_{j,l}\right)^2}{K-1}} \qquad \text{(III-2)}$$

Condition $j,\ j \in \{0,1,...\}$,

Listener $k,\ k \in \{1,2,...,K\}$,

Talker $l,\ l \in \{1,2,...,L\}$ *(L is equivalent with the number of files per test condition),*

with $MOSLQS_{j,l}$ as "Mean Opinion Score Listening Quality Subjective" for talker *l* of condition *j* as defined in:

$$MOSLQS_{j,l} = \frac{1}{K}\sum_{k=1}^{K} v_{j,l,k} \qquad \text{(III-3]}$$

III.2 The standard deviation for condition-based analysis

The standard deviation σ_j for condition-based analysis is defined as follows making use also of the definitions in the previous subclause:

$$\sigma_j = \sqrt{\frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\left(v_{j,l,k} - MOSLQS_{j,l}\right)^2}{N-1}} = \sqrt{\frac{K-1}{N-1}\sum_{l=1}^{L}\sigma_{j,l}{}^2} \qquad \text{(III-4)}$$

N denotes the number of votes per condition.

III.3 Exceptional cases

The calculation of σ_j and $\sigma_{j,l}$ as well the corresponding ci95 as described above is the regular mode of calculation. However, it requires access to the individual votes or at least to a per-file standard deviation and/or confidence interval along with the number of votes.

For some of the existing databases, this information is not available. In these cases, the following simplifications will be applied:

1) If only the *MOS* per-condition is provided, then this value is used for the per-file evaluation as well. This per-condition MOS will be used as MOS for each file. The required ci95 per-file is obtained according the following rules in that case.

2) If only *standard deviation* per-condition is provided, then this value is used for the per-condition evaluation instead of σ_j as described above. The systematic over-/under-prediction of a speaker or a file may influence (increase) that value.

3) The ci95 per-file (required for secondary analysis) can be derived by the simplification that the standard deviation is constant and equal for all files in that condition. In this simplification the ci95 per-file can be derived by:

$$ci_{95}(per\ file) = t(0.05, M)\frac{\sigma(per\_condition)}{\sqrt{M}} \qquad \text{(III-5)}$$

with M as the number of votes per-file.[2]

4) If only the *confidence interval ci95* per-condition is provided, then this value is used for the per-condition evaluation directly. The systematic over-/under-prediction of a speaker or a file may influence (increase) that value.

5) The ci95 per-file (required for secondary analysis) can be derived again by the simplification that the standard deviation is constant and equal for all files in that condition. In this simplification the ci95 per-file can be derived by:

$$ci_{95}(per\_file) = ci_{95}(per\_condition)\frac{\sqrt{N}}{\sqrt{M}} \qquad \text{(III-6)}$$

with N as number of vote for the entire condition and M as the number of votes per-file.1

6) If neither *confidence interval* nor *confidence interval* is available the ci95 (per-condition) is fixed as 0.2. The ci95 (per-file) is calculated as in 2) above.

[2] In case that M is unknown, it might be assumed by N divided by the number of files scored in one condition (i.e., if N = 96 and for files were scored in one condition, M = 24).

Appendix IV

Normality test

(This appendix does not form an integral part of this Recommendation.)

The normality of the vote distributions for each file can be evaluated by calculating the skewness and kurtosis coefficients of the analysed distribution. The associated standard error of skewness (SES) and standard error of kurtosis (SEK), defined as follows:

$$SES\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

$$SEK = 2SES\sqrt{\frac{n^2-1}{(n-3)(n+5)}}$$

These indicate if skewness and kurtosis measures deviate from expected values for normal distributions. If their values are less than –2 or greater than 2 then it is very likely (95 per cent confidence) that the distribution has excessive skewness or kurtosis.

To determine if the considered distribution deviates are statistically significant from a normal distribution, it is necessary to combine the kurtosis and skewness tests. The d'Agostino-Pearson omnibus χ^2 test can be used for this purpose:

$$K^2 = Z_{g1}{}^2 + Z_{g2}{}^2$$

Where $Z_{g1} = \frac{SKEW}{SES}$ and $Z_{g2} = \frac{KURT}{SEK}$ represent the Fisher z statistics of the statistical significance test.

If K^2 is greater than χ^2(2) = 5.991 then the distribution is considered as not normal, with 95 per cent confidence.

Appendix V

Statistical significance of the rmse_tot* across all experiments

(This appendix does not form an integral part of this Recommendation.)

Another metric based on the absolute prediction error, a value *rmse_tot** can be calculated for each tested algorithm.

$$rmse\_tot_v{*} = \frac{1}{K}\sum_{K} rmse{**}_{k,v}$$

with

$$rmse{**}_{k,v} = \sqrt{\frac{1}{N_k - d}\sum_{N_k}\left(\frac{MOSLQS(i)_k - MOSLQO(i)_{k,v}}{\max(ci_{95}(i)_k\,, c)}\right)^2}$$

where the index *i* denotes the condition or the sample in experiment *k*; N_k determines the number of test conditions or number of samples in that experiment respectively; K denotes the number of <u>all</u> considered experiments; and d the number of degree of freedom. The degree of freedom *d* is set to 4 when a 3rd order mapping is applied prior to the values. The constant *c* avoids divisions by very small *ci95* values as they usually appear at the low end of the scale. The value of *c* is set to 0.1.

To determine if two or more algorithms are statistically equivalent or not, a statistical significance test can be applied to the aggregated distance measures $rmse\_tot{*}_v$ available for all models, where $rmse\_tot{*}_{min}$ is the lowest $rmse\_tot{*}_v$ in the evaluation. *t* describes the degree of freedom that is the total number of samples (files or conditions) in <u>all</u> considered experiments.

$$r_v = \max(0, \frac{rmse\_tot\,{*}_v^2}{\left(rmse\_tot_{\min}^2 + c\right)} - \mathrm{F}(0.05, T\ \ , T\ \ ))$$

If $r_v = 0$, then the model *v* is considered statistically equivalent to the model with $rmse\_tot{*} = rmse\_tot{*}_{min}$. When $r_v > 0$, then the model *v* is considered statistically significantly worse than the lowest $rmse\_tot{*} = rmse\_tot{*}_{min}$.

Bibliography

[b-ITU-T Handbook] ITU-T Handbook (2011), *Practical Procedures for Subjective Testing*.

[b-ITU-T P-Sup.23] Recommendation ITU-T P-series – Supplement 23 (1998), *ITU-T coded-speech database*.

[b-Benjamini-1] Benjamini, Y. and D. Yekutieli, *The Control of the False Discovery Rate in Multiple Testing under Dependency*. The Annals of Statistics, 2001. 29(4): p. 1165-1188.

[b-Benjamini-2] Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society. Series B (Methodological), 1995. 57(1): p. 289-300.

[b-Berger] Berger, Jens (2011), *About MOS and Quality Measurements, White paper, SwissQual AG, Zuchwil, Switzerland*.

[b-Brunnström] Brunnström, K. and M. Barkowsky, *Statistical quality of experience analysis: on planning the sample size and statistical significance testing.* Journal of Electronic Imaging, 2018. **27**(5): p. 11.

[b-Holm] Holm, S., *A Simple Sequentially Rejective Multiple Test Procedure*. Scandinavian Journal of Statistics, 1979. 6(2): p. 65-70.

[b-Maxwell] Maxwell, S.E. and H.D. Delaney, *Designing experiments and analyzing data: a model comparison perspective*. 2nd ed. 2003, Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc.

[b-Spiegel] Spiegel, M. (1998), *Theory and problems of statistics, McGraw Hill*.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2020