

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.1401

(07/2012)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

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



International
Telecommunication
Union

Documents
Document Sharing Hub

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400

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

Copyright © 2012 ITU-T. All rights reserved.

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
1.0	ITU-T P.1401	2012-07-14	12

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/>.

© ITU 2013

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

	Page
1 Scope	1
2 References.....	1
3 Definitions	2
4 Abbreviations and acronyms	2
5 Conventions	2
6 Subjective test and objective algorithms	2
6.1 Aspects related to subjective testing.....	2
6.2 Aspects related to objective algorithms.....	3
7 Evaluation framework	4
7.1 Data preparation	4
7.2 Analysis types.....	5
7.3 Prediction on a numerical quality scale.....	6
7.4 Uncertainty of subjective results	14
7.5 Statistical evaluation metrics	14
7.6 Statistical significance evaluation	17
7.7 Statistical evaluation in the context of subjective uncertainty: epsilon insensitive rmse and its statistical significance	19
7.8 Statistical evaluation of the overall performance	21
8 Guidance on algorithm selection	22
8.1 Per experiment performance.....	23
8.2 Overall figure of merit.....	23
8.3 Worst performance cases.....	23
8.4 Averaging statistical metrics across experiments.....	23
9 Special cases	23
9.1 Evaluation of algorithms with more than one output	23
9.2 Evaluation of algorithms against pre-defined minimum performance requirements	23
10 Demonstration cases	24
Appendix I – Algorithm mapping to the subjective scale.....	27
Appendix II – The impact of the third order versus first order mapping.....	29
II.1 Application of third order and first order mappings.....	29
II.2 Gain of third order mapping	29
Appendix III – Confidence intervals calculation	31
III.1 The standard deviation for file-based analysis	31
III.2 The standard deviation for condition-based analysis	31
III.3 Exceptional cases.....	32
Appendix IV – Normality test.....	33



	Page
Appendix V – Statistical significance of the rmse_tot* across all experiments	34
Bibliography.....	35

Document released under license

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 (2011), *The E-model: a computational model for use in transmission planning*.
- [ITU-T G.1070] Recommendation ITU-T G.1070 (2012), *Opinion model for video-telephony applications*.
- [ITU-T J.247] Recommendation ITU-T J.247 (2009), *Objective perceptual multimedia video quality measurement in the presence of a full reference*.
- [ITU-T J.341] Recommendation ITU-T J.341 (2011), *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 (2011), *Perceptual objective listening quality assessment*.

- [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-T P-Sup.23] Recommendation ITU-T P-series – Supplement 23 (1998), *ITU-T coded-speech database*.
- [ITU-R BS.1116] Recommendation ITU-R BS.1116 (1997), *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*.
- [ITU-R BT.500] Recommendation ITU-R BT.500 (2012), *Methodology for the subjective assessment of the quality of television pictures*.

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 [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).

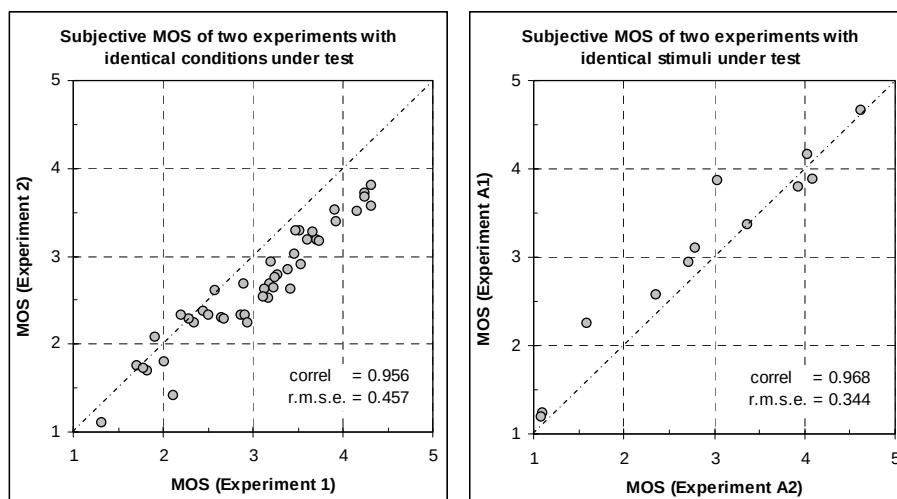


Figure 7-1 – Comparison of MOS values obtained in different subjective experiments

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.