SwissQual

A Rohde & Schwarz Company

An Introduction to Measuring Quality and MOS Testing

Dr. Jens Berger

October 2011

SwissQual AG
Allmendweg 8 CH-4528 Zuchwil Switzerland
t +41 32 686 65 65 **f** +41 32 686 65 66 **e** info@swissqual.com
www.swissqual.com

Part Number: 23-000-201127-2

SwissQual has made every effort to ensure that eventual instructions contained in the document are adequate and free of errors and omissions. SwissQual will, if necessary, explain issues which may not be covered by the documents. SwissQual's liability for any errors in the documents is limited to the correction of errors and the aforementioned advisory services.

Confidential materials.

All information in this document is regarded as commercial valuable, protected and privileged intellectual property, and is provided under the terms of existing Non-Disclosure Agreements or as commercial-in-confidence material.

When you refer to a SwissQual technology or product, you must acknowledge the respective text or logo trademark somewhere in your text.

SwissQual®, Seven.Five®, SQuad®, QualiPoc®, NetQual®, VQuad®, Diversity® as well as the following logos are registered trademarks of SwissQual AG.

Diversity Explorer™, Diversity Ranger™, Diversity Unattended™, NiNA+™, NiNA™, NQAgent™, NQComm™, NQDI™, NQTM™, NQView™, NQWeb™, QPControl™, QPView™, QualiPoc Freerider™, QualiPoc iQ™, QualiPoc Mobile™, QualiPoc Static™, QualiWatch-M™, QualiWatch-S™, SystemInspector™, TestManager™, VMon™, VQuad-HD™ are trademarks of SwissQual AG.

SwissQual acknowledges the following trademarks for company names and products:

Adobe®, Adobe Acrobat®, and Adobe Postscript® are trademarks of Adobe Systems Incorporated.

Apple is a trademark of Apple Computer, Inc.

DIMENSION®, LATITUDE®, and OPTIPLEX® are registered trademarks of Dell Inc.

ELEKTROBIT® is a registered trademark of Elektrobit Group Plc.

Google® is a registered trademark of Google Inc.

i.Scan is a trademark of CommScope, Inc.

Intel®, Intel Itanium®, Intel Pentium®, and Intel Xeon™ are trademarks or registered trademarks of Intel Corporation.

INTERNET EXPLORER®, SMARTPHONE®, TABLET® are registered trademarks of Microsoft Corporation.

Java™ is a U.S. trademark of Sun Microsystems, Inc.

Linux® is a registered trademark of Linus Torvalds.

Microsoft®, Microsoft Windows®, Microsoft Windows NT®, and Windows Vista® are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries U.S.

NOKIA® is a registered trademark of Nokia Corporation.

Oracle® is a registered US trademark of Oracle Corporation, Redwood City, California.

SAMSUNG® is a registered trademark of Samsung Corporation.

SIERRA WIRELESS® is a registered trademark of Sierra Wireless, Inc.

TRIMBLE® is a registered trademark of Trimble Navigation Limited.

U-BLOX® is a registered trademark of u-blox Holding AG.

UNIX® is a registered trademark of The Open Group.

Contents

Figures

Tables

1 Measuring Quality

"Beauty in things exists merely in the mind which contemplates them."

David Hume's Essays, Moral and Political, 1742

It is difficult to have an absolute measure for "quality". It is not the same as trying to measure a physical quantity such as weight, length, or brightness.

To produce a valid measure of "quality" you need to take into account the perceptions, cultural attitudes, preferences, and expectations of human beings, which can vary in any group of people. Quality perception is driven by experience and expectation.

Essentially, quality is the gap between what you expect and what you actually receive.

A "value" or a "score" for quality depends on the quality expectation of the person you ask, that is, quality measurement is to some extent a moving target. However, unlike measures such as speech intelligibility or listening effort, quality measures take the user expectation into account. This approach makes quality measures just as important as the other measures for the optimization of services. A provider is not interested in achieving the highest possible data throughput or the most transparent speech codec that is technically feasible, but rather is interested in a system performance that matches the customer expectation in the best possible way.

How do you measure quality?

Everyone has a different perception of "quality". For meaningful results, you need to ask a statistically significant number of people to judge and to assign quality scores to samples of speech, music, video, game playing or whatever you are trying to measure. These samples are experienced by the human subjects under identical and tightly controlled conditions.

People can also become habituated to certain levels of "quality". For example, when mobile phone technology was first deployed people found the audio characteristics of the codec difficult to listen to. Over time they "tuned in" to the characteristics of the codec and began to find the quality more acceptable.

The quality scores that are derived in this way are only valid for the specific conditions in which the tests were conducted and for the specific questions that the human subjects were asked. In order to produce a truly useful measure of "quality", the conditions should reflect real life as closely as possible.

2 Subjective and Perceptual Experiments

The traditional way to measure quality is a subjective experiment in which people are placed in a simulated environment in a laboratory, for example a telephone environment with a telephone device and a handset for listening.

The advantage of such an environment is that you can keep the many influences on the perceived quality constant for the participants over the course of the experiment. For example, for a telephone listening quality test, the environment would be the same room, the same handset, and the same voice stimuli for each participant.

In addition to testing and measuring quality experiments, there are also perceptual experiments, which you can use for other topics such as speech intelligibility or listening effort. These experiments use the same principles and similar setups to the subjective tests.

Experimental Setup

In principle, you can freely define the setup for a subjective quality experiment based on the object that you want to test and the focus of your investigation.

An experiment can be focused on many aspects, for example, telecommunication or can be focused on the interaction between speech dialogue systems, or, switching channels in TV services. You can also use customer surveys in which people have assigned scores directly after a telephone call or after using a data service in their home environment or in a laboratory.

An experiment in a lab environment allows more control over the test setup, which can be kept constant for each person in the test. However, there is a wide range in the degree of control as well, for example, in a conversational test. Such a test can be influenced solely by the people who are having the conversation, for example, their knowledge and competence in the language, individual pronunciations, strategies to solve a problem in a conversation and of course their respective empathy for one another.

On the other side, there are very efficient tests for listening quality or visual quality where a person assigns a score to a short stimulus of a few seconds in length. Since the experiments are simulating a pure listening or viewing situation, no interaction by the person is required[1]. Here the speech excerpts are pre-recorded and can be selected for content, phonological balance, and neutral pronunciation.

Such tests allow the assessment of a large amount of test stimuli in a short time within a controlled lab environment. In principle, the more controlled the test and scoring environment is the more reliable and reproducible the scores are. Furthermore, fewer people have to be involved to obtain statistically reliable data. On the other hand, an experiment in a laboratory remains artificial and the design and setup must be chosen very carefully. A slightly misbalanced setup can lead to biased results even if the results appear to be statistically confident.

In each case, test subjects must be familiar with the test situation, for example, each person should have had experience in making a phone call or in watching TV. That is, you need to invite people who already have a quality expectation from using that kind of service in daily life. However, these people should not have existing knowledge of the voice or video samples that you use in order to avoid training effects on a certain sample. To ensure that people do not experience these effects they are usually only invited every six months or so to participate in the individual test situation. Depending on the test case, groups of individuals can be invited, for example, young people to test a youth portal or elderly people to evaluate special voice or video services aimed at this group. However, common tests that address common telephony or video services are conducted in a balanced distribution of gender, age and educational status.

Laboratories for voice and video tests have to fulfil a large set of strict requirements with respect to room acoustics, lighting, test devices, and the quality of the presentation equipment.

[1] Listening Quality refers to quality that is perceived in a pure listening situation. No own speech activity required as for example, in conversational tests. The person just listens and scores. The same applies to Visual Quality; here a pure viewing situation is simulated.

Test Types and Scales

The typical subjective tests for listening or visual quality can be divided into two main categories: Tests in which a person compares a stimulus with a reference or undistorted source signal, and tests in which the person cannot directly compare the received signal to the undistorted reference signal.

Subjective tests can be rated by the following methods:

- **CCR (Comparison Category Rating):** Samples A and B are compared to each other and scored as better or worse with labels such as "much better" or "slightly worse". The scale has a negative and a positive boundary with the mid-point equal to zero.
- **DCR (Degradation Category Rating):** Sample B is compared to a reference sample A and scored as equal or worse with labels such "degradation slightly distorting". The scale ranges downward from a high score to a lower score, but with positive values.
- **ACR (Absolute Category Rating):** One sample is presented for scoring. The scoring labels reflect absolute terms such as "good" or "bad". The scale ranges downward from a high score to a lower score, but with positive values.

There are many variations and derivations of these basic test types. There are tests in DCR design where the reference test sample stimulus pair is presented twice in series or even multiple times on request by the test person. Another approach is called MUSHRA (for audio / voice) where all test samples can be listened to on request through a selection panel and are then scored relative to each other. The person can re-listen to a sample as often as he or she wants and in any order he or she wants. The person can even switch to another sample during the playback of a sample.

The CCR and DCR methods have many individual test procedures during which the reference and the test stimulus are presented sequentially or even multiple times. Special procedures allow a person to switch between both stimuli at any time and to repeat the presentation. This approach allows one to fine tune a codec algorithm, the selection processes, and so on. These procedures also provide a fine grading of differences and distortions with respect to an unprocessed signal.

However, to get an impression of the real end user quality experience, it is often not helpful to compare to the undistorted reference stimulus directly. In a real environment, for example, during a phone call, listening to a CD or MP3 or watching TV, a person never has the possibility to compare their sample to an "original". Instead, the person compares the experience to his or her expectation, which is driven by own experience. A person has an expectation of how a telephone or an instrument sounds and what a clear picture on TV looks like.

The Absolute Category Rating (ACR) tests are used to simulate such situations and do not involve a direct comparison to a reference stimulus.[2] In such a test the person must rate the quality on an absolute scale based on his or her own expectation and experience. The most common rating method for an ACR test is a following five point scale, which has verbal categories in the native language of the test subjects.

[2] Although there is no direct comparison to the undistorted reference signal, this reference signal is presented at any place in the experiment without indication to the listener or viewer. This approach is often called ‘with hidden reference’.

Table 2-1 Five point quality scale

Score	English	German	French	Spanish
5	excellent	ausgezeichnet	excellente	excelente
4	good	gut	bonne	buena
3	fair	ordentlich	assez bonne	regular
2	poor	dürftig	médiocre	mediocre
1	bad	schlecht	mauvaise	mala

Each test subject provides a quality score for a stimulus or a test condition, for example, a pre-recorded speech or video sample of a few seconds in length. The experiment yields a collection of individual scores that is independent from the scale that was used. These score can be a collection of even integer values, if you use 'buttons' in the experiment for scoring, or real numbers, if you use a continuous scale and score with a slider device or a computer mouse.

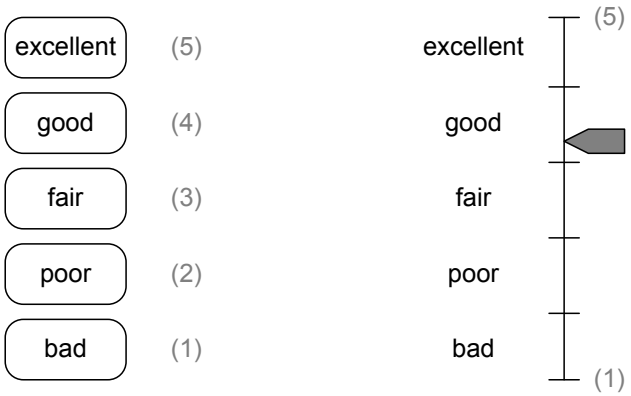

Figure 2-1: Examples of five-point scales as used in ACR experiments

The example ACR scales in Figure 2-1 are typically used in ITU-T and ETSI activities. For special tasks other scales and labels are used, for example, a seven-point or an eleven-point scale. These scales often use the score as a number instead of verbal labels.

Note: In this paper, all MOS examples are obtained with and refer to five-point MOS scales, where five corresponds to "excellent" and one to "bad".

Mean Opinion Score

The so-called "Mean Opinion Score" (MOS) is the average of the individual scores for a given stimulus or test condition. The resulting MOS value indicates the quality of the short sample. Traditionally a MOS is derived by carrying out listening or visual tests with groups of human subjects that are large enough to constitute a statistically significant sample.

The term MOS is only a generic definition and is meaningless without a further specification of the kind of quality perception that the MOS describes. A MOS can be obtained for listening quality as well as for visual quality through different test setups, scales, labels, and questions asked[3].

[3] In technical applications the quality measurement is performed to characterize a transmission or processing system. Simplified, the quality of a certain speech or video sample is interpreted as the quality of the system under test. However, the MOS takes into

Each subject's score is, as previously discussed, driven by his or her individual experiences, expectations, and preferences. In practice there is always a variation in the scores that each person assigns to a sample. The score is also subject to the short-term nature of the test as well as unintentionally assigned incorrect scores. Consequently, the MOS is the average of a distribution of individual scores.

However, in practice each individual in a statistically significant group of people will assign different values to a sample, even if the sample is undistorted. Some subjects will lack confidence in their own perception, some will be hypercritical, and still more will award a less than perfect score purely through accident or because of mental distraction during the test. As a result, the highest MOS value reached in subjective tests is around 4.5.

At the lower end of the quality scale we have a corresponding but slightly different effect. This difference is caused by the fact that the lower end of the "quality" scale is much broader than the higher quality end, that is, the score can be 'worse than bad'. Unlike the lower quality end of the scale, the upper end cannot be so easily extended, that is, the quality cannot be "better" than speech or video which has not been distorted in any way.

To constitute statistical significance, a group of subjects should consist of at least 24 people. In scientific papers MOS values are accompanied by their standard deviation to provide some basic information about the width of distribution of the individual scores. An additional value that is often given along with the MOS is the 95% confidence interval. This interval is a statistical range around the MOS where the true MOS of the whole population will fall with 95% probability and provides an impression of how close the MOS is to the 'true quality'. Logically, this confidence interval becomes smaller when the group of subjects and the number of votes in the test increases. In a well-designed traditional test this range of uncertainty is less than 0.2 MOS.

However, even an efficiently planned and conducted subjective ACR experiment using stimuli of a few seconds remains limited to approximately 200 stimuli. This amount can be assessed in approximately one hour, which is the usual maximum duration of an experiment, including pauses.

Therefore, test conditions and scores of different experiments are often compared, especially if no experiment has been conducted to cover all of the conditions of interest 'at once'. The following section discusses the relation of scores in one experiment and between different experiments.

Basics of Test Design

Within an experiment, a set of different stimuli or test conditions need to be scored. Condition refers to a given transmission condition, such as codec under a defined bit-rate or a defined type of a noise reduction system with a certain setting or similar during which the processing components do not change. For each condition a set of voice or video files is usually processed. However, the quality of samples that are processed under the same condition might vary slightly as the effect of the processing on each sample is always slightly different.

For this reason, a set of different samples is processed and scored. In case the scores become averaged across all scores that were obtained for the same test condition, the MOS describes the quality to be achieved with that condition in a more general sense. This approach minimizes content dependencies of the MOS value of a test condition.

A listening quality test usually uses a set of four different voice samples for each condition. The samples are spoken by male and female talkers. To avoid training effects for each condition, a different set of samples, that is, other spoken text, is used or a variation in the used texts. For visual quality tests, the dependency on the content is much higher and different examples of typical content categories are used. The selected samples can be reused for each condition in video tests.

account the content, it describes what the user perceives, not what the processing system does. To interpret a MOS as a qualitative description of the transmission system requirements on content have to be matched.

The following terms are important for understanding test designs.

- **Non fractional design:** Everyone in the test listens to or views the same test samples so that a sample that has been processed by one condition is presented to everyone in the test. As a result, the number of scores for the sample is equal to the number of people who participated in the experiment.
- **Partially fractional design:** The participants are divided into groups of six or eight people. Each group listens to different samples and each person in the same group listens to the same sample. This design minimizes the content dependency by increasing the variation of source samples, that is, texts or talkers. The disadvantage of this design is that each sample is scored by fewer people than in other tests, which in turn results in a lower statistical confidence. The partially fractional design makes the most sense if the goal of the subjective experiment is a 'per condition' MOS.
- **Full fractional design:** Each person listens to or views different samples for each condition. Essentially this design is the same as the partially fractional design, except that each group only consists of one person. In other words, only one score exists for each sample. Due to the high number of samples involved, the number of votes per condition is the same as the other designs.

The following table illustrates these different designs. A typical test has 50 conditions and 24 test subjects. Each person listens to four samples of each condition, that is, two male and two female voices.

	Non fractional	Partially fractional	Full fractional
Number of conditions	**50**	**50**	**50**
Number of test persons	**24**	**24**	**24**
Samples listened per condition and person	**4**	**4**	**4**
sub-groups	1	4	24
persons per 'sub-group'	24	6	1
Votes per sample	24	6	1
Votes per condition	96	96	96
Samples per condition	4	16	96
Total samples in experiment	200	800	4800

Table 2-2 Examples for different fractional designs of subjective experiments

Non-fractional designs used to be very common for speech coding standardization efforts. Recently such designs have been superseded by fractional designs, especially full fractional designs when a certain content dependency of conditions is expected. However, fractional designs are only useful when a 'per condition' MOS is targeted. In such a case, a 'per-sample' analysis is statistically unconfident due to the small number of scores for each sample. In fractional designs, the number of voice samples to be processed is much higher.

Each approach has its pros and cons. In the latest standardization activities for objective models, non-fractional or partially fractional designs were applied. For visual tests the amount of data is higher by sizes when compared to voice tests and hence non-fractional design is used. In recent codec standardization activities, full fractional designs have been used.

It is not enough to define an arbitrary set of conditions and then process the test samples. To design an effective experiment we must also consider the so-called context effects. These effects are the result of the scoring behaviour of a test person who has been influenced by his or her daily experience with telecommunications, which can lead to changing expectations over the long term. An example of an expectation change is the deployment of mostly noise-free ISDN networks throughout Germany in the 1990s. After this deployment, in contrast to telephony tests that were conducted before this deployment, a noise-free sample was expected and even slight noises received low scores. Conversely, the introduction of mobile phones made people familiar with coding distortions from the EFR or AMR speech codecs. Such distortions have now become tolerated in listening experiments. These codecs now receive relatively higher scores in subjective experiments than when they were first introduced because people have become familiar with the distortions. The same effect is also expected in situations that involve Voice Over IP or synthetic voices in automated voice services and dialogue systems. The long term effects of this phenomenon can lead to

different experimental scores for the same distortions over a longer observation period.

The general design of a subjective experiment leads to intra-experiment effects in a 'medium term' manner. These effects are caused by the stimuli that are presented in the test as well as short term intra-experiment context dependencies that are caused by the presentation order.

An example of a short term effect is a medium quality voice sample that receives an artificially higher quality score because the sample follows a low quality sample in the presentation order. Conversely, if the medium quality sample follows a high quality sample, the medium quality sample receives an artificially lower. Short term effects also occur when different distortion types are presented; however the effects tend to be more complex. These effects can be efficiently minimized by creating individual presentation orders for each person or for each group of people. That is, each person listens to the same conditions and stimuli, but in a different presentation order.

People tend to rate samples relative to the overall context of the experiment, which results in a 'medium term' context effects. These effect are responsible for most of the differences in intra-experiment scoring results. For example, in an experiment that only uses very high quality samples, a person might struggle with his or her expectation of bad quality. Instead, the person tends to score medium quality samples lower than expected. Other context effects can arise during an experiment due to the mixture of test conditions that cover different distortion types. For example, in an experiment that only consists of noise-free voice samples and one or two noisy samples, the latter are immediately seen as totally different from the main corpus. The noisy samples usually score much lower when compared to an experiment with a balanced distribution of noise-free and noisy samples.

Before an experiment starts, participants listen to or watch a selection of training samples that cover the range of distortions and qualities that are present in the test. This approach helps each person to fix develop his or her expectation of bad or good quality within the context of the experiment.

In addition, and more importantly, a well-balanced design is very important to minimize context effects. In a balanced design the selection of conditions and stimuli covers the entire range of the topic to be studied as well as the entire quality range.

Furthermore, so-called reference and anchor conditions are often included in the experiments, for example, emulated distortions such as noises or filtering. Emulations have the advantage of scalability and exact reproducibility. For example the amount of added noise can be adjusted in exact steps of the signal-to-noise ratio. Anchor conditions, that is, standardized coding algorithms can also be adjusted.

Results that are obtained from a subjective experiment depend on many influences. Due to smart and sophisticated design constraints the negative effects can be minimized but never avoided completely. In addition, a MOS value is simply an average of a limited number of individual votes and remains a subject of uncertainty and depends on the group of people in the experiment. The following section discusses the comparability of MOS values that have been obtained in different experiments.

Comparing MOS Values of Different Experiments

The group of people within an experiment remains the same and the stimuli, that is, the 'context' covers the same range of test conditions. Even though accidental individual false scorings can occur, the relation of the MOS in this experiment can be considered as ‘true’, but only for this group and this test design. Individual scores also drive the confidence interval, which only describes the uncertainty of the MOS values in this experiment.

Only MOS values from similar tests 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 from a listening test that uses an ACR scale cannot be directly compared to MOS values from a DCR experiment. However, when MOS values from the same test types are compared, some limitations are present. The results of each test are slightly different even if the same participants are involved.

Scores assigned by a listener or viewer are never the same in a repeated experiment that presents the same samples in the same order. This phenomenon can be imagined as a type of noise that overlays the MOS scores.

Subjects are also influenced by the short term history of an experiment, that is, the samples they scored directly before the current one. An example of a short term context dependency is when subjects score a

medium quality sample higher if it follows one or two low quality samples. Similarly, if a medium quality sample follows very good quality samples, subjects tend to score the medium quality sample lower. A strategy to average out this short term context dependency is to use different presentation orders for each participant; however, the statistical uncertainty remains.

The largest difference between subjective experiments is caused by medium and long term context effects. The medium term context effects are the set of conditions for the experiment, that is, the average quality, the distribution of qualities in the experiment in general, and the occurrence of individual distortions. For example, in an experiment that contains mainly low quality samples, subjects tend to score such samples higher and vice versa. This tendency is due to participants using the entire range of the quality scale in an experiment without regard to the verbal category labels. As a result, subjects tend to adapt the scale to the qualities that are presented in the experiment. In addition, individual distortions that only occur in a few samples during an experiment are scored lower than when the distortions occur more frequently in an experiment due to the increased familiarity of the subjects to the distortions. This familiarity is a mid-term context dependency that reflects effects that have been caused by the design of the individual experiment.

Experiments also have long term dependencies, which reflect the general cultural behaviour of the individuals, for example, the exact interpretation of the category labels, cultural attitude of quality, language dependencies, as well as daily experience with telecommunication or media. Here we also have to face the fact that quality experience and expectation may change over time. People become familiar with the quality of mobile codecs and the associated distortion until they simply accept the distortion as part of their mobile phone experience. Furthermore, the same telecommunication channels are relatively noiseless compared to POTS telephony decades ago.

All these effects lead to differences between individual experiments. The effects cannot be avoided, but can be minimized through informed instruction, a well-balanced test design, a sufficient number of participants, and a mixed presentation orders for the samples.

Such influences can lead to different MOS values for the same test condition or 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 design 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 compare MOS values from different experiments directly.

In the past, MOS values from individual experiments were transformed to a technical scale called Quantization Distortion Unit (qdu). This transformation was achieved through the introduction of reference conditions into each experiment to simulate distortions as produced by logarithmic PCMs. This approach worked out well so long as the distortions of the test conditions were similar in sound. However, qdu-based alignment became impractical with the rise of CELP and other coding algorithms as well as tests of other distortions such as PCM quantization noises.

The following diagrams illustrate this problem. Within an ITU-T activity, a set of 44 different test conditions (focussing on G.729 in 1995) was defined. Over these test conditions voice samples in different languages were transmitted and scored in the individual laboratories by native subjects for each language. Even though the conditions, the test design, and the listening conditions were exactly the same, there are differences in the MOS values. The left graph in Figure 2-2 shows the results that were obtained in a Canadian laboratory with North American samples (Experiment 1) on the x-axis and the same experiment conducted in Japan with Japanese samples and listeners on the y-axis (Experiment 2).

Note: How do you interpret a scatter-plot?

A common way of presenting scores of experiments or of subjective and objective scores to each other is a so-called scatter-plot. Each point represents one condition or test sample, where two values are available, that is, from two different analysis methods. One score is defining the value on the x-axis the other score is on the y-axis.

The better the two methods or data sets match, the more narrow the plotted points are distributed along the 45° line in this diagram (like a 'pearl-chain').

In the example of the left diagram, the points fall a bit below the 45° line. That means that the MOS of Experiment 2 shown at the y-axis are a bit lower than those ones of Experiment 1. A point above that line gives an indication that the method shown on the y-axis gives a higher (more optimistic) score. In principal any kind of narrow distribution ('pearl chain') shows that the rank-order can be widely reproduces, even the interpretation of the scale may different. Besides general trends, scatter plots are excellent illustrations of the width of distribution, which gives an indication how consistent the results of the two compared methods are. Even more, individual outliers (as in the right diagram) can be detected easily.

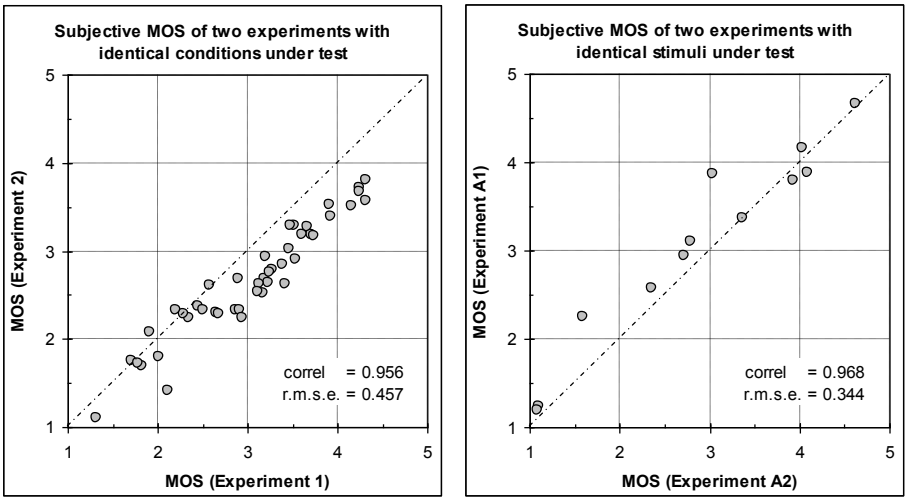

Figure 2-2: Comparison of MOS values obtained in different subjective experiments

If the MOS values for each condition were the same, all points would be on the 45° line; however, the points are mostly below the line. This means 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.[4]

Of course, such effects can easily be declared as language dependent, however this is not entirely the case. Considerable differences are also present when experiments were conducted in the same laboratory with identical stimuli, for example, the results on the right graph in Figure 2-2. Both experiments were conducted in the same laboratory with the same test equipment and source samples, that is, texts and speakers, and the focuses of the experiments were very similar. For formal reasons the group of the 24 test people was different.

Both experiments use identical voice stimuli as anchor conditions to provide an overlapping area to illustrate the differences. In principle, the qualitative ranking of the stimuli is the same in either experiment. However, there is a bias in the lower area and some larger differences in MOS even though the stimuli are identical.

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

Both examples show 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 in an experiment, which can influence the 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 in 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 range. In such a case, people tend to use the whole scale for the range of quality that you include.
- **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. The assumption is that the relative quality ranking is constant and can always be reproduced.

  In practice, such a ranking cannot always be reproduced by another subjective test.

  Firstly, 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 a range of 0.15 MOS, no finer resolution than 0.3 MOS in ranking can be achieved with a certain confidence.

  Besides statistical uncertainties, another problem can lead to real changes in the relative quality ranking of systems. We assume that a subjective experiment can always determine the quality ranking of different systems correctly even when the MOS values contain statistical uncertainties. The MOS values are always correct but depend on the design of the experiment, for example, the distribution of distortion types throughout the experiment. An under representation of a distortion leads to a relatively low MOS value while an over representation of the distortion leads to a high MOS value as participants become desensitized to the distortion. Such context dependent effects can 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.

However, in the end a subjective experiment remains a closed set. The experiment is only true to itself and can hardly be compared with other experiments.

3 Sample and Content Dependencies

Traditionally, artificial signals were used for measurements and analysis, for example, test videos with circles, grey steps, lines, and coloured squares. In analogue or waveform-based telecommunications systems sweeps, that is, sine waves scrolling through spectrum or noises were used.

Test signals needed to cover the entire range of the 'width' to be transmitted. For example, there were tests for Telex services that transmitted the entire alphabet, the digits, and the control commands in sequence. Similarly, when voice samples started to be used as test signals the test samples would include sentences in which each letter in the alphabet occurred at least once.

In modern quality analysis the focus is on using real contents as test signals in the same manner as these contents are used in a service. The goal is to create short test sequences with the same characteristics as averaged arbitrary contents and which represent the main characteristics of a language or a genre. These sequences form quasi archetypes of a certain application.

This approach became very important with the introduction of source coding in which characteristics of the source signal are modelled. The model parameters are transmitted and composed in the receiver. Modern speech coding algorithms follow this approach with the application of a model of the human vocal tract. Sine waves and noises are not suited for this type of testing as only voice samples can represent the phonetic distribution of a spoken language. Similar strategies are applied to video where objects in the images are recognized, parametrically described, transmitted, and redrawn. Artificial contents can over or under represent objects or spatial complexity.

In the on-going effort to standardize test signals the ITU-T and ETSI have specified some test signals for dedicated measurements, including voice samples. The same VQEG and ITU-R has a selection of video clips that are often used in quality evaluations. However, these clips should be seen as examples as it does not help to standardize a single or small set of voice or video clips for testing. Compression algorithms and measurement methods could be optimized for such clips and may even show excellent performance, but still worse in ‘real world’ applications.

Content Requirements

A participant in a quality test has no knowledge about the transmission or processing systems under test. He or she is simply asked to assess the quality of the sample that is presented. Consequently, the perceived quality depends on the degradations that are introduced by the system under test, for example, a transmission channel, the source sample, and interactions between that specific sample and the transmission system.

The goal is to minimise the bias of the quality values due to the content to obtain quality values that accurately characterise the system under test. To help to achieve this goal, the test content should meet the following criteria:

- Represents the service under test, for example, human voice for telephony, head and shoulder video for video telephony.
- Free of distortions or characteristics which can be interpreted as distortions, for example, video clips with special effects such as blinking objects, stretched faces, and so on
- Free of highly emotional content
- Natural source, for example, a voice recording of a native speaker in a quiet studio, normal video content such as people and landscapes
- Familiar to test participant, that is, voice sample is in his or her native language, video sample shows common scenes in his or her country, and so on

These criteria can only lower content effects. To accurately characterise a transmission system, the content must be carefully chosen. For voice samples, the spoken texts should reflect the phonological characteristics of the language and in a short sample if possible. In the video domain samples of typical genres should be defined and used. Samples of more than one type of content are often required to obtain a range of different quality levels or to perform averaging over the quality values for a test condition or a content type.

Content Dependency and Cultural Behaviour

If a system is tested with the same subjective quality test but in different laboratories, slight differences emerge in the quality scores.

Differences in voice test scores of different languages are usually due to factors such as one language having a significantly higher occurrence of voiced parts than a language that consists of many unvoiced parts. Such languages would be processed differently by an encoder.

However, it is typically the individual structure of the sample that causes this difference. Certain imperceptible characteristics that are present in the original unprocessed sample might affect certain transmission systems, for example, thresholds for pause muting.

Perhaps the main reason for the differences in the scores for different languages is the cultural attitude of the test group. This attitude might result in small differences in the interpretation of the quality labels, that is, different experience in daily life and therefore different expectations of quality as well as culturally driven 'scaling'. Some cultural groups may be more generous and will assign good scores as long as there is no major degradation in quality; while other cultural groups who are not as generous who will assign good scores only for absolutely perfect quality.

Differences in quality scores that are caused by context and cultural effects cannot be reproduced by an objective predictor. However, as previously mentioned, the qualitative rank order normally remains unaffected and the predictor will rank the sample in the same order in which they would be ranked by a subjective test.

4 Always a Question of Context

Telephony Speech versus Wide Band Speech

Telephony speech has a typical sound. The limited audio bandwidth has been accepted for decades. When listening quality is scored in a telephone context, the subjects use a traditional handset and all samples reflect this band limitation. Consequently, the listener adapts his or her internal quality scale to the environment, which means that a perfect but band limited transmission is the best he or she will expect. As a result, a score of 4.5 is expected.

However, the situation is quite different for wide band audio. During the listening test a participant is confronted with wide band signals that obviously sound better than narrow band speech. The quality expectation is now more driven by the audio and hi-fi experience of the user, which can even be forced with hi-fi headphones. This experience and expectation will lead to an adaption of the internal scale as well. Higher scores such as 'excellent' are now reserved for undistorted wide band speech. In contrast, clean telephony speech with band limitation will be scored lower even if the signals yield excellent quality results in a pure narrowband situation. This example shows that the test scenario and context has an extreme influence in the interpretation of MOS values.

Telecom industries are currently initiating the evolution from narrowband telephony to wideband speech transmission. The wideband codecs are ready and have been approved by the standardization bodies while the handsets have not been restricted in processing power and the core networks are being upgraded.

Logically, the test and evaluation methods for voice telephony must be adapted to a wideband scenario. As a consequence, the known and long-time accepted MOS values for narrowband channels will be dropped. A perfectly designed narrowband channel might be scored at least 0.5 MOS lower in a wideband context as such a sample is clearly inferior to wideband speech.

TV versus HD and QCIF versus QVGA

When television was introduced a few decades ago, the number of lines was restricted to a range of 500 to 600. This resolution continues to set our quality expectation of television and even when a test uses high quality samples, the resulting MOS values for these samples is 4.5.

High Definition TV (HDTV) is becoming increasingly popular, especially due to the LCD flat panels and the move to digital transmission. Similar to the wideband telecom scenario, HDTV sets a higher expectation of image quality than standard TV. Perceptual tests that includes Standard TV (SDTV) and HDTV samples usually digitally up-scale the SDTV samples for an HDTV screen. As a result, the participants view the quality of such samples as degraded since SDTV signals contain much less information than HDTV signals. The same perception can also occur between two HDTV signals of different resolutions, that is, 720 lines or 1080 lines. To summarize, a perfect SDTV signal will always be rated lower in an HDTV context since the expected reference for high quality is based on the HDTV experience.

Similar effects can be observed on smart phones that support QVGA (320x240 pixels) or 480x360 pixels. These phones usually automatically up-scale samples to the native display resolution. Mobile DVB-H TV content is transmitted in QVGA resolution, which means the video clips are already in this resolution.

Note: YouTube provides streams in all popular resolutions; QVGA (320x240 pixels) is named 240p, 480x360 as 360p or VGA (640x480) as 480p.

Streaming of perfectly encoded and un-disturbed content in the common QCIF format (144 x 176 pixels) will be scored lower on such a smartphone screen. Due to up-scaling, a QCIF video appears as un-sharp and blurry since this format contains much less image information than the QVGA format. This effect becomes even more drastic on smart phones with WVGA resolution, which often involves 800x480 pixels.

To summarize, it is important to consider the context of the underlying quality when interpreting the results of a video test, for example, a SDTV signal in a HDTV context will receive a lower quality score than an HDTV.

5 Obtaining Objective Measures for Quality

Principal Approaches for Objective Quality Prediction and Estimation

The basic principle of an objective measure or estimation of quality is to predict a perceived quality by means of technical analysis.

Such an analysis can simply be the knowledge about the architecture of the system under test (glass box). These models are based on experienced quality values for individual components and construct an overall quality through the sophisticated combination of the individual parameters. Such models are often called 'planning tools', for example, the E-model, which is recommended by ITU-T as G.107.

A grey box is another approach for a system under test where some information has been derived through technical analysis. For example, you can try to predict the quality of an IPTV application or VOIP call by analysing the IP-stream.

However, the prediction accuracy remains quite limited since the actual voice or video medium is not analysed. Furthermore, only degradations that are visible in the observed link can be estimated. For example, a quality estimation that is based on the RF parameters in a mobile connection does not consider a degradation that has been inserted into the core network or at the other end of the connection. Similarly, analysing the IP link to an IPTV customer does not consider degradation at the TV source or at the head-end. Algorithms that only analyse the IP stream for quality prediction are often called a bit stream model.

For the most reliable quality prediction, you need to analyse the media signal, which is perceived by the user and which can be subjectively scored subjectively as well. The signal can be a WAV file for voice or audio or a non- or de-compressed video file, for example, in 24bit RGB format. In such a content based approach, the system or the network under test can be handled as black box without knowledge of the components, architecture, or design outlines.

However, and especially for IP video, there have been investigations to develop so-called hybrid models, which combine a content analysis with information that has been derived from the bit-stream to improve the prediction accuracy.

The following block schemes illustrate the content base, bit-stream, and hybrid approaches with respect to IP video testing.

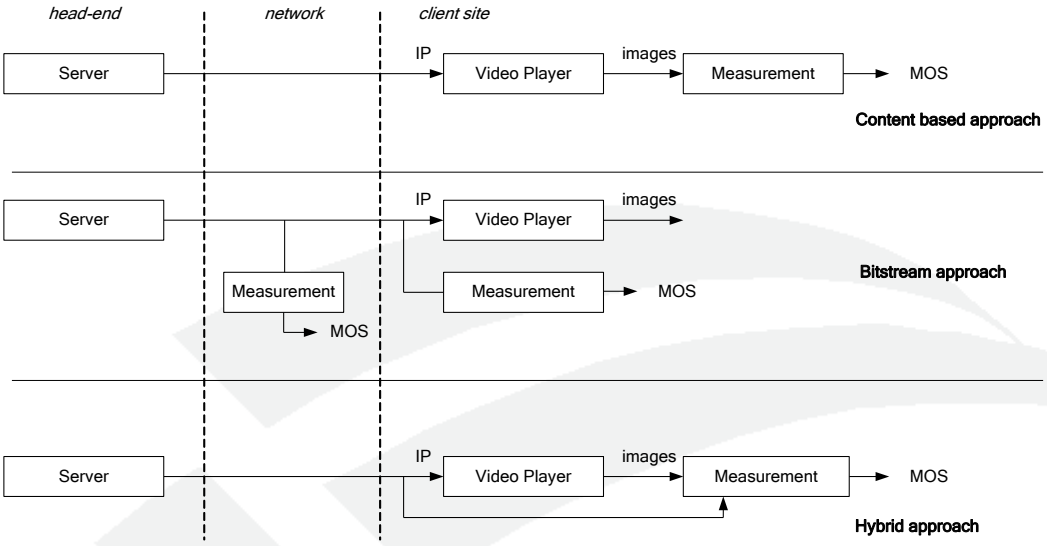

Figure 5-1: Basic scheme of content, bit-stream and hybrid approaches

Huge improvements have been made to content based approaches over the last two decades. Traditional signal analysis, that is, signal to noise ratio analysis, have been replaced with algorithms that model human perception. For voice and audio analysis, human hearing is emulated with psycho-acoustic approaches while human voice is covered by cognitive models of speech perception. The psycho-visual models for video are

less evolved; however, modern algorithms combine traditional signal analysis with psycho-visual ideas.

Content based approaches have the advantage that they analyse the sequence as it is seen by a user. The sequence covers the artefacts that were collected in the transmission until the point of presentation. In contrast, a bit-stream approach can only consider artefacts in the observed IP link. Artefacts that are introduced prior to and after the link are ignored. Combining content based and bit-stream methods into a hybrid model can take advantage of the two approaches. Currently there are no hybrid models on the market. However, they will probably first become available for IPTV and multimedia TV solutions where pure content based models cannot achieve the accuracy of voice or audio models.

Accuracy and Statistical Evaluation of Objective Measures

In principle, any objective measure or estimator of quality tries to predict a quality as it is perceived by a group of users. Therefore, objective measures for quality are different from the traditional empirical approach for measuring physical phenomena. Objective measures simply estimate or predict quality as it would be perceived by a large group of human observers. To develop and calibrate objective measures such as VQuad, VMon, or SQuad, a large amount (many thousands) of subjectively scored samples are required. The new ITU-T Recommendation P.863 POLQA was developed and evaluated with more than 45000 subjectively scored speech samples.

Objective approaches do not measure quality in the traditional sense and instead try to predict the score that might be obtained from a sufficient amount of people. However, some uncertainty remains since there is no true MOS for a certain stimulus or condition that can be used for training. A MOS, even for the same stimulus, can differ between individual experiments for the training process of an objective model. This problem cannot be solved as 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 5-2 shows the test results for four experiments under the exact same conditions (speech codecs, frame and bit errors, and background noises) and which have been carried out in four different languages and laboratories.

Note: These test results are publically available from the ITU-T as Supplement 23.

The left diagram in Figure 5-2 shows the MOS values of subjective experiments. The MOS values for American English are plotted over the x-axis while the value for French, Italian, and Japanese are on the y-axis. The latter three experiments are plotted versus the American English results.

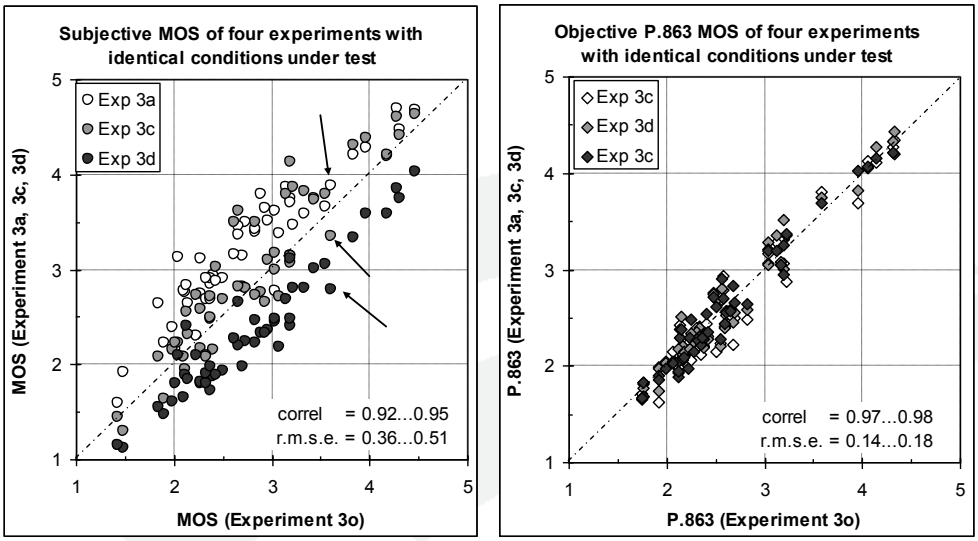

Figure 5-2 Subjective and objective scores of experiments covering the same test conditions

The three arrows in the left diagram represent a single test condition in all of the experiments. The MOS in the American English experiment (on the x-axis) is about 3.6 while the MOS for the other experiments (on

the y-axis) are at 2.8 (Exp. 3d), 3.4 (Exp. 3c), and 3.9 (Exp. 3a). Each experiment has a narrow distribution in itself; however the scores of each experiment have different offsets. The large spread indicates considerable differences between the experiments.

The right diagram shows results that were derived from the same experiments with objective quality predictions from the new ITU-T P.863 POLQA instead of MOS values. It is obvious 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. As only the processing conditions were identical and not the test samples, small differences between the results remain.

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

Figure 5-3 shows the objective prediction scores are plotted versus the MOS values for experiments 3c and 3d.

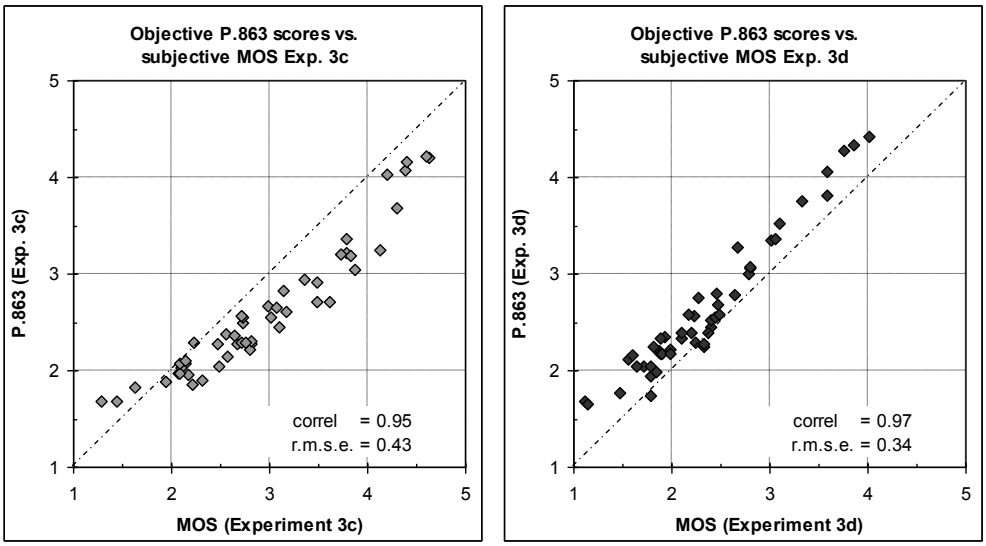

Figure 5-3: Objective vs. subjective scores of two experiments of ITU-T Suppl. 23

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° line. These offsets are not a problem or a malfunction of the objective model, but are rather caused by the inter-experiment differences of the subjective experiments. The objective model predicts a kind of an 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 key question is whether an objective model accurately predicts the quality as an averaged MOS across many subjective experiments or an individual subjective test that is driven by many influences.

In Figure 5-2 and Figure 5-3, the diagrams use statistical measures to show the performance or accuracy of the objective measures. More specifically, Pearson's correlation coefficients and root mean square errors (r.m.s.e.) are plotted on the diagrams.

The difference between the apparently true MOS and the outcome of the objective measure is often referred to as a prediction error. However, a more accurate way 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.

There are different ways to minimize the inter-experiment differences before calculating a prediction error. The traditional qdu method does not work with modern test setups. To assess the performance of an objective method, it is important to show how well the qualitative rank-order can be reproduced. This means that inter-experimental effects that do not change the rank order, such as biases or individual gradients, can

be removed before comparison to an objective method.

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 r.m.s.e. without changing the rank-order and by the optimal function of a given structure, that is, first and third order polynomials. Mapping function examples can be found in the related P.862, P.862.1, P.563, P.863, and J.341 ITU-T Recommendations. 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.

The following figure illustrates the effect of mapping on the data set from experiment 3c.

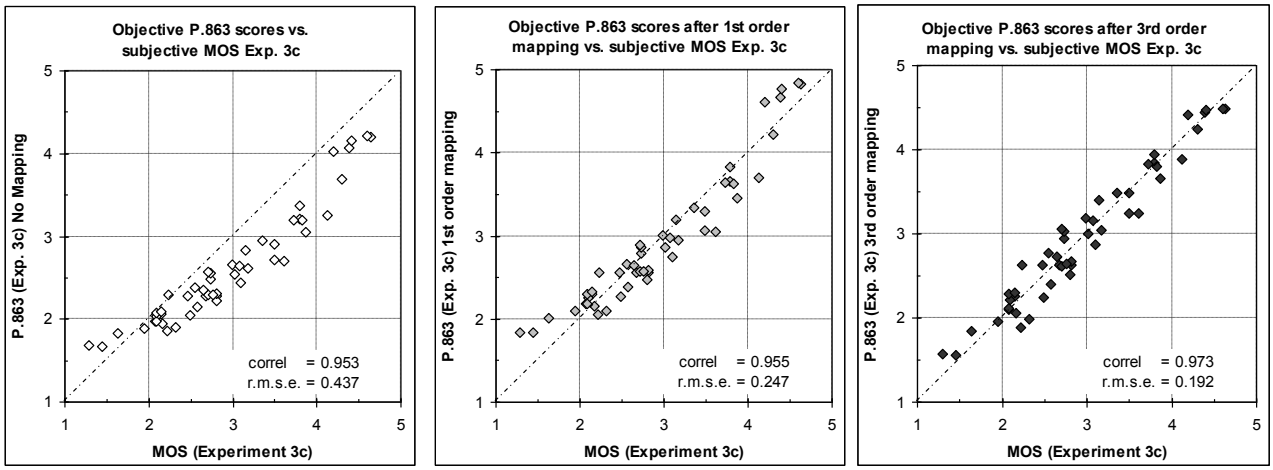

Figure 5-4: Objective vs. subjective scores a raw scores, after 1st order and after 3rd order mapping

The left diagram gives an impression of the relation between the objective raw scores, that is, ITU-T and P.863, and the subjective MOS values. There is a clear offset, a slightly different gradient, and the data form a banana shape.

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

$$y' = a + by, \text{ where } r.m.s.e.(x, y') \rightarrow min$$

As a result, the offset is compensated and the gradient is closer to the 45° line. Consequently, the r.m.s.e. 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 right scatterplot shows the results after applying a third order polynomial function. The equation has the form:

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

This mapping corrects the offset and the gradient and linearizes the banana shape. Due to the improved linearization, the r.m.s.e. drops to 0.19 for this experiment while the Pearson correlation coefficient goes up to 0.97.

Note: The rank order of the scores has not been changed. The part of the third order polynomial that was used was chosen by monotonous constraints.

In summary, modern best objective approaches are reaching an accuracy that is close to the uncertainty of the MOS values in well-designed experiments and can often fall below the differences between individual experiments.

Objective Measures as Models of Subjective Experiments

This section will attempt to describe what an objective measure is actually predicting. In principle, an objective measure predicts the subjective scores that are used for its training process, that is, the measure

predicts an average across the experiments. The term subjective experiment covers a wide range of potential experimental setups and it makes no sense to train a model with data that has been obtained in totally different setups. The MOS values are not comparable and have even a different meaning depending on the setup that was used. At the very least, the experiments that are used for training should be based on the same principle that is, listening only tests should use the five-point ACR scale. Furthermore, the test context should be the same, that is, narrowband telephony or video on an HDTV screen.

Under this logical pre-assumption, an objective model predicts the MOS values that would be expected in the experiment setups that were used to train the model. The ITU P.862, P.563 and P.863 Recommendations are all models of listening only tests in narrowband telephony that use five-point ACR scales.

The following quotation is from ITU-T Recommendation P.563:

> *"… It only measures the effects of one-way speech distortion and noise on speech quality in the same way as it can be investigated by an auditory test assessing listening quality on an ACR scale. The P.563 algorithm scores the speech signal in that way, as it is presented to a human listener by using a conventional shaped handset and listening with a SPL of 79 dB at the ERP. "*[5]

On the other hand, there is a dilemma. Even this description of the experimental setup appears quite restricted and in practice there is still a wide range of interpretation of this setup. This again leads to remaining differences between the individual subjective experiments as seen in Figure 2-1. If the objective measure is well-designed and falls into the accuracy range of these inter-experiment differences, it cannot become any better by using more training data.

The only way to achieve better prediction models is through the increased quality of the experiments that are used to train the models. This can only be reached by stricter design rules and an increased number of participants in each experiment. On the other hand, this limits the scope of the objective model further. The model will predict a very accurate MOS which can only be derived by tests following these design constraints.

Full-Reference versus No-Reference Assessment

If you want to manage the quality of your mobile network, you need to be able to accurately assess the quality. One method to assess the service quality of a telecommunications network is to determine the quality of a signal that is transmitted through the network.

In the case of objective quality evaluation, several approaches are available to assess this quality. The primary distinction in objective quality evaluation is between a no reference method, that is often called non-intrusive or single-ended, and a full reference method, that is often called ‘intrusive’ or double-ended approaches.

- **No reference:** Evaluation and rating is only conducted and based on the received signal. Examples of this single-ended method are a test call to an answering machine or even live monitoring.
- **Full reference:** A reference signal is transmitted and the received signal is evaluated and rated based on the known reference. This double-ended method requires a test call.

Both methods predict the Mean Opinion Score (MOS), the score that would be obtained by performing a subjective test. The basic relationship between subjective and objective assessments and full and no reference models is shown in Figure 5-5.

[5] ERP: Ear Reference Point; SPL: Sound Pressure Level

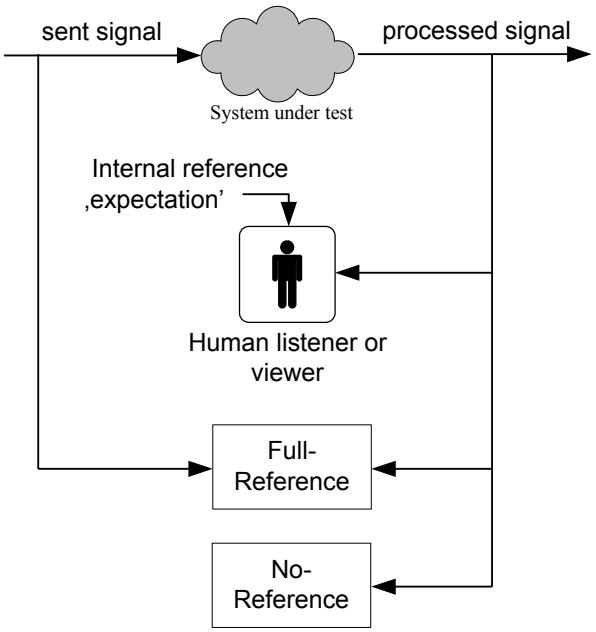

Figure 5-5 Subjective versus objective quality assessment

Usually, the accuracy of a no reference approach is lower than that of a full reference approach due to the lack of reference signal for a detailed comparison. However, the accuracy may be sufficient for a basic classification of the quality measure and for the detection of consistently poor quality links. No reference models have a wider application range because there is no need for special facilities at the far end of the communications link under test.

Application of Full and No Reference Quality Measures

Full reference or no reference only describes if the algorithm uses a reference signal for comparison or not. A better distinction is made with respect to the intrusiveness of the measurement. For a measurement, an intrusive measurement requires a dedicated test channel, that is, a network resource. For example, the measurement might use all of the measurement devices in a test system for dedicated test calls for quality measurement. Non-intrusive measurements do not interfere with the network at all. For objective quality testing, the two basic methods can be used in several application scenarios.

Full Reference Double-Ended Measurements

In a full reference double-ended measurement, both ends of the connection are under control and a defined voice, data, or video sequence is transmitted over the test connection. The setup requires a controlled answering station or server with known stored sequences at the far end side. Logically, a test connection must be established and resources must be taken from the network. A disadvantage of this approach is the need to intervene at the source of the signal and the network under test. However, the advantage is that the input signal or reference signal is known, which allows for the very accurate and detailed analysis of voice or video quality impairments. By applying models of human perception, each change in the signal during transmission can be detected and evaluated for the impact on perceived quality. The full reference methods are also applicable to optimisation processes in laboratories as well as in real networks. The methods are capable of measuring even minimal degradations of the signal and can be applied to compare various transmission scenarios.

There are two basic applications. The first application is where the input and the output of the system under test are connected to the same measurement unit in a laboratory environment. The unit can simultaneously send and record the signal.

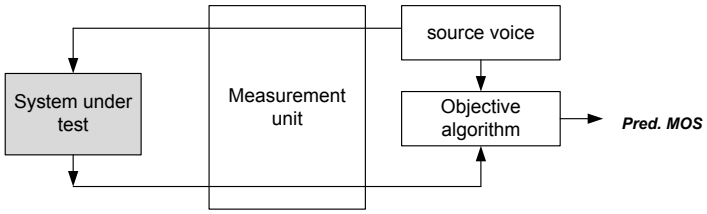

Figure 5-6: Full-Reference test approach for local applications

The source signal, in this example a voice signal, is available in the measurement unit and can be used to feed the system under test or as reference signal.

However, for measurements out in the field this approach does not work. The input and the output of the system are usually in different locations. Instead, a modified approach in such a situation can be applied for full reference testing.

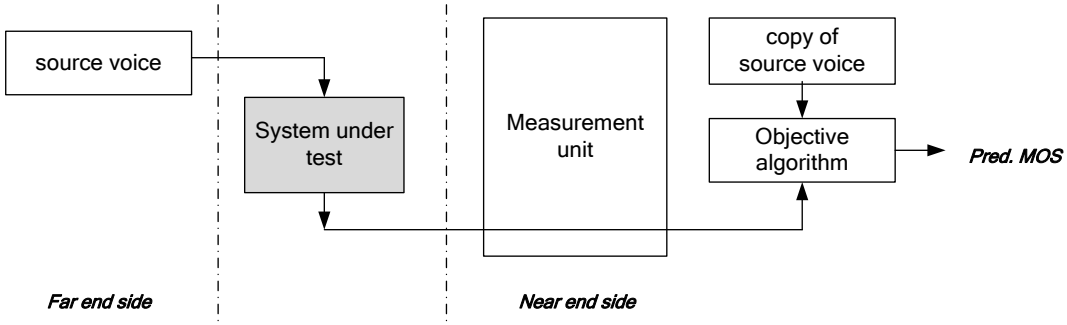

Figure 5-7: Full-Reference test approach for distributed application

It has to be ensured that at the far end side (feeding the system under test) the source signal is stored and played out on request. The measurement unit at the local near end side records the signal and hand the recording over to the objective prediction algorithm. The algorithm also requires a copy of the source signal on the local side. This approach is common practice for all types of field and network measurement systems. Some wrapping components are used as synchronization frameworks for synchronized play-out and recording as well as for maintaining the far end answering station.

No Reference Single Ended Measurement

A test connection is established to an answering station, which plays an unknown voice or video signal to the receiving side, for example, from a streaming server or a live TV application.

Note: This test connection places an additional load on the network resources under test.

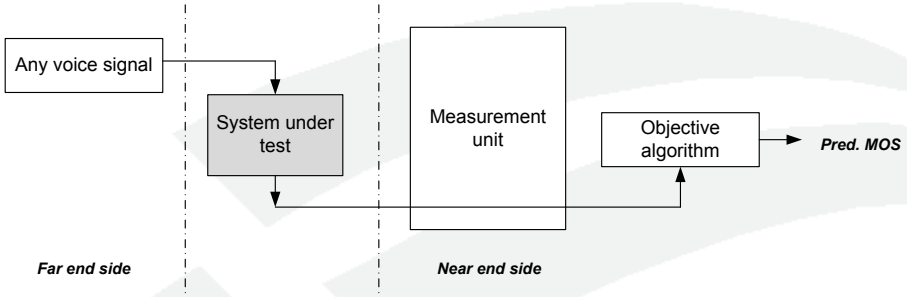

Figure 5-8: No Reference but intrusive test approach

As the no reference approach does not require information about the source signal, any answering station that plays a voice signal can be used, for example, a weather forecast. However, the influence of the quality of the source signal has to be taken into account in the final quality score, for example, a noisy weather forecast is scored lower due to the noise in the source signal.

No Reference Non-Intrusive In-Service Monitoring

An example of non-intrusive in-service monitoring is the assessment of video signals in real applications such as IPTV or telephony by parallel monitoring the active connections in the core network without a known reference signal. For this scenario, a dedicated test connection is not required and a no reference method running on the equipment of friendly users as they go about their everyday business can be used. However, this means that the measurement point can be anywhere in the network. Any degradation beyond this measurement point is not unconsidered. In addition, the characteristics of the specific device being used can affect the measurement.

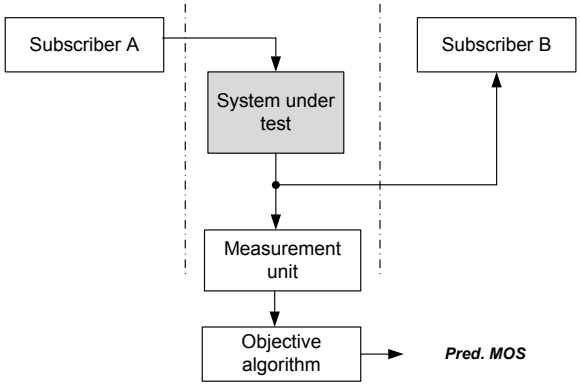

Figure 5-9 No-Reference and non-intrusive test approach for monitoring

In-Service Monitoring can also be used for a telephone call between two persons during which the quality is qualitatively scored. Since a dedicated test connection is not required, the test can be called non-intrusive. Usually, these measurements are performed in the network close to the end with the user's interface, which is also called the mid-point. Measurements that are performed directly at the user's side are called end-point measurements.

In principle, all of these measurement methods can also be used to record stimuli, that is, voice samples, in subjective experiments. However, the large amount of data can only be efficiently handled by objective algorithms running in real time on the measurement unit. A typical approach is to record and to store the stimuli for spot checks later on. There are even procedures in place to reduce the amount of data by only keeping the recorded stimuli in which the objective algorithm or another component in the measurement unit has indicated problems.

6 Conclusion

Quality measurements and the obtained results depend on many experimental factors and objectives. The related ITU-T P.800, P.830, and P.910 ITU-T recommendations only define only a basic framework for subjective tests. Within this framework a very wide range of individually designed experimental setups are possible without violating the constraints of the standard. This framework allows for a wide application area of the basic test approaches as well as individual experiments. However, the latter cannot be compared directly. For a correct interpretation of the results, additional information to the test setup and design is required.

Objective predictors can model subjective test procedures and can predict subjective results with some uncertainties. Object quality predictions can either use a reference signal for comparison or no reference signal. These basic approaches open or limit the measurement area and the practicality of the objective measurements as well as increase or decrease the accuracy of the quality predictions.

Modern cutting edge psycho-acoustic motivated measurements can already achieve highly accurate quality assessments. Uncertainties tend to not be much larger than the statistical confidence intervals of a subjective test. Other methods are less accurate but much faster or easier to implement, for example, for network monitoring.

Practical use cases for objective quality measurements are troubleshooting or quality monitoring where a lower accuracy is sufficient for spotting severe and systematic problems. Other applications might require more accuracy, for example, system optimization. In such a situation, more complex models and a wider range of test signals are required.

Objective quality measurements are efficient tools for quality estimation. These tools avoid some of the shortcomings of subjective experiments, but cannot achieve greater accuracy than the experiments that were used for their development and training process.