International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.863.1

(05/2013)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Methods for objective and subjective assessment of speech quality

---

Application guide for Recommendation ITU-T P.863

Recommendation ITU-T P.863.1

ITU-T

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.

Recommendation ITU-T P.863.1

Application guide for Recommendation ITU-T P.863

Summary

Recommendation ITU-T P.863.1 provides important remarks that should be taken into account in the objective quality evaluation of speech conforming to Recommendation ITU-T P.863. Users of ITU-T P.863 should understand and follow the guidance given in this Recommendation.

This Recommendation is a supplementary guide for users of Recommendation ITU-T P.863, which describes a means of estimating listening speech quality by using reference and degraded speech samples. The scope of Recommendation ITU-T P.863 is clearly defined in itself. This Recommendation does not extend or narrow that scope; rather, it provides necessary and important information for obtaining stable, reliable, and meaningful objective measurement results in practice.

History

Edition	Recommendation	Approval	Study Group
1.0	ITU-T P.863.1	2013-05-14	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/.

Table of Contents

Page

Recommendation ITU-T P.863.1

Application guide for Recommendation ITU-T P.863

1 Scope

This Recommendation is a supplementary guide for users of [ITU-T P.863], which describes a means of estimating listening speech quality by using reference and degraded speech samples. It assumes that an objective quality algorithm strictly conforms to [ITU-T P.863]. This can be confirmed by the conformance test provided in [b-ITU-T P.Imp863].

This Recommendation does not extend or narrow the scope of [ITU-T P.863]; rather, it provides necessary and important information for obtaining stable, reliable and meaningful objective measurement results in practice. It also provides example results for common situations and explanations for how certain degradations impact the score.

Additional characterization results will be added to this Recommendation, in the form of appendices, as this information becomes available.

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 P.10] Recommendation ITU-T P.10/G.100 (2006), *Vocabulary for performance and quality of service*; Amendment 2 (2008); Amendment 3 (2011).

[ITU-T P.863] Recommendation ITU-T P.863 (2011), *Perceptual Objective Listening Quality Assessment.*

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the terms and definitions defined in [ITU-T P.10].

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

AGC Automatic Gain Control

ASL Active Speech Level

IRS Intermediate Reference System

MIRS Modified Intermediate Reference System

MOS	Mean Opinion Score
MOS-LQO	Mean Opinion Score – Listening Quality Objective
NB	Narrow Band
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
SWB	Super Wideband
VoIP	Voice over IP

5 Conventions

None.

6 Introduction to ITU-T P.863

[ITU-T P.863] provides an objective speech quality measurement algorithm for measuring the voice quality of narrowband, wideband and superwideband networks.

6.1 History of objective speech quality

Speech quality has no physical definition. Inherently, people just have an opinion about when something sounds good or bad. However, stakeholders need to be able to quantify the quality delivered by a telephone system in order to maximize investment and ensure adequate service is provided to customers. For many years, the only effective manner by which to determine the quality of a telephone network was to perform a subjective test. Subject testing, explained more in the next clause, involves asking a panel of users what they think of a recording or connection. The panel typically vote on a five point scale, and the average of the votes is deemed to be the quality of the connection. This number is called the mean opinion score (MOS).

The running of subjective tests is time consuming and costly. During the late 1980s compression technologies were introduced in digital networks to increase capacity while reducing costs. Before their introduction, it was generally possible to determine the performance of a network using simple tone-based measurements. With the introduction of new speech processing technologies, it was found that results from tone-based techniques could contradict users' experiences. A new measurement methodology was required. The increased availability of general purpose computing allowed the development of computer programs capable of modelling the results of subjective tests.

In 1996, Recommendation ITU-T P.861 (perceptual speech quality measure (PSQM)) [b-ITU-T P.861] was published. The core concept introduced in this first generation algorithm was that human hearing could be modelled to extract a representation of audible differences between a reference and a degraded pair of signals, and that these differences could be mapped to the scores of subjective tests.

Shortly after [b-ITU-T P.861] was published, work was started to address practical limitations of the first generation model in terms of its applicability for testing networks. This work led to the publishing of a significantly improved model called perceptual evaluation of speech quality (PESQ), which was published as [b-ITU-T P.862] in 2001, together with the withdrawal of [b-ITU-T P.861]. Work continued on [b-ITU-T P.862] for a number of years, for example, with the introduction of a wideband extension in 2005. However, as more complex signal processing was added to the telephone network, it became clear that a new model was required.

In 2006 ITU-T initiated a new activity for the development of the third generation model. The intention was to provide a backward-compatible model that could also assess new speech signal processing technologies as well as the anticipated move to superwideband networks. The result of this work was published as [ITU-T P.863] at the beginning of 2011. It should be noted that the introduction of [ITU-T P.863] does not deprecate [b-ITU-T P.862].

6.2 Basics of a subjective test

It is important to consider subjective testing when discussing objective speech quality assessment. Essentially, subjective testing underlies all three generations of objective models described above. The intention for this type of objective model is to predict the result of a listening quality subjective test for a given reference and degraded file pair.

ITU provides various Recommendations for performing subjective tests; the most relevant for [ITU-T P.863] are [b-ITU-T P.800] and [b-ITU-T P.830].

A subjective test aims to find the average user opinion of a system's speech quality by asking a panel of users a directed question and providing a limited-response choice. For example, to determine listening quality, users are asked to rate 'the quality of the speech' by selecting from the scale shown in Table 1.

Table 1 – MOS

Opinion	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

A MOS is calculated for a particular condition by averaging the votes of all subjects. This type of test is described as an absolute category rating (ACR) experiment.

The results of a subjective test are influenced by many factors and a great deal of effort is made, in the planning of the subjective test, to control these factors. Regardless of the planning effort, a score for a given condition from a subjective test is not a definitive result; run the same test with another set of subjects and a different result will be received. The expected range of results for a given condition can be expressed as a confidence interval. Some subjective tests provide results with a smaller confidence interval than others. Place the same condition in another subjective test, with a different quality bias, and it will more than likely be found that the score compared to the first test systematically moves up or down. There is a tendency to see subjective tests maintain rank ordering of conditions, but not absolute scores.

It is difficult to cope with these inconsistencies. Thus, it is not uncommon for companies to claim that a particular coding technology is superior to another technology and provide subjective test results to support these claims, while others companies are able to demonstrate a contrary conclusion.

Over the past two decades, the importance of normalizing subjective scores onto a 'common' scale, for the purpose of developing and evaluating objective measurement algorithms, has absorbed significant effort. [b-ITU-T P.1401] describes this process in detail. Later in this Recommendation, a procedure is presented for removing simple subjective-test condition biases using linear-regression; this step represents a minimum level of normalization that should be applied before comparing subjective and objective results.

6.3 Benefit of using Recommendation ITU-T P.863

The two main reasons for using an ITU recommended objective speech quality algorithm are:

– the ability to compare results between different organizations;

– model development represents a small part of the effort; validation over many thousands of conditions represents the bigger effort.

6.4 Relationship with other Recommendations

- [b-ITU-T P.862] – predecessor to [ITU-T P.863] for objectively measuring listening speech quality of narrowband and wideband networks. This Recommendation has not been deprecated by the introduction of [ITU-T P.863].
- [b-ITU-T P.563] – a no-reference, single-ended, objective listening speech quality method for predicting the subjective quality of narrowband networks.
- [b-ITU-T G.107] (E-model) – a planning tool that combines various network performance parameters to determine a quality score on a 100 point scale. This score is referred to as the R-factor.
- [b-ITU-R BS.1387] – an objective tool for measuring perceived audio quality.
- [b-ITU-T P.800.1] – a Recommendation defining MOS terminology. Based on the definitions in this Recommendation, [ITU-T P.863] scores should be labelled as mean opinion score – listening quality objective (MOS-LQO) if presented alongside subjective or estimated scores.

6.5 Moving from ITU-T P.862 to ITU-T P.863

[ITU-T P.863] narrowband mode was developed to provide backward compatibility with [b-ITU-T P.862]. This backward compatibility means that [ITU-T P.863] can be used wherever [b-ITU-T P.862] is currently being used. It also means that the predictions of [ITU-T P.863] can be directly compared to both new and old narrowband subjective tests.

Although built to be backward compatible, different scores will be seen when testing a connection with [ITU-T P.863] in place of [b-ITU-T P.862]. Different scores will be seen because there are fundamental differences between how [b-ITU-T P.862] and [ITU-T P.863] process the signals. In addition, a significantly larger database of subjective test results has been used in the development and calibration of [ITU-T P.863] as compared to [b-ITU-T P.862]. This increased data provides a more complete view of how people judge the full range of speech quality delivered by modern telephone networks. For this reason, and because [ITU-T P.863] achieves better performance across all databases compared to [b-ITU-T P.862], predictions from [ITU-T P.863] should be viewed as more accurate than [b-ITU-T P.862].

It is recommended that [b-ITU-T P.862] and [ITU-T P.863] be run in parallel until an appropriate feeling is developed for [ITU-T P.863] scores. Additional information on the differences between [b-ITU-T P.862] and [ITU-T P.863] are the topic of clause 12.

6.6 Challenges with ITU-T P.863

The biggest challenge of using [ITU-T P.863] is in selecting appropriate reference material. [ITU-T P.863] does not always produce a perfect score when a reference file is compared with itself. This is described as a *reference transparency* issue.

The transparency issue arises because [ITU-T P.863] assumes that a reference signal will have a balanced timbre and it judges deviations from a balanced timbre as degradation. If the timbre in the reference is not balanced, for example, because of a bass-boost or a lot of sibilance, a reference-reference comparison with no impairments will not return the maximum score for the scale. Typically, one will see a drop of a 0.1 or 0.2 MOS; however, it can be larger.

In addition to ensuring reference transparency, further constraints have been placed on the reference's content structure to ensure that reported scores are representative of user perception. These additional constraints have been added following practical experience because [b-ITU-T P.862] has been misused over the years.

It should be noted that [ITU-T P.863] does not enforce any reference content constraints, thus it is still possible to use non-conforming reference material. This may produce interesting results in internal studies, but values should not be reported publically if the reference signal does not follow the constraints.

6.7 MOS misconceptions

A number of minor misconceptions exist around objective listening speech quality assessment and what a MOS value represents. Some misconceptions are:

- The effect of a constant end-to-end delay is not assessed by subjective or objective listening quality test methods. The presence of delay variation may affect the score, particularly during periods of active speech.
- The effect of talker echo and echo delay are not assessed by subjective or objective listening test methods. A listener cannot hear talker echo.
- Objective listening test methods cannot assess mixed signals, such as in the double talk condition where there is sidetone at the listener end, or multiple concurrent speakers on a conference bridge.
- MOS is not only a measure of voice quality. MOS is the acronym of Mean Opinion Score. A MOS is the result of any subjective test where subjects are asked to vote on a scale (discrete or continuous) and the votes are averaged across subjects to determine the mean opinion of the subjects. It is equally valid to have a MOS prediction for video and audiovisual content as it is for speech content.

7 Operational modes

[ITU-T P.863] provides two operational modes:

- Narrowband: predicts quality as it would be perceived through an intermediate reference system (IRS) type receive filtered monaural handset.
- Superwideband: predicts the quality as it would be perceived through a diffuse field equalized Hi-Fi headphone for diotic (same signal in both ears) listening.

[ITU-T P.863] always applies an IRS receive filter to the reference and degraded input signals in narrowband mode. This is not the case in superwideband mode. [ITU-T P.863], regardless of mode, predicts the perceived quality of a recording on a five point ACR scale.

7.1 Why two operational modes?

Telephone networks contain a mix of audio-bandwidths, including traditional 300-3400 Hz narrowband, wideband up to 7 kHz and superwideband from 50-14 kHz. Until now, most telephone networks have provided only narrowband connections and users' quality expectations are biased for this bandwidth-range. Adding wider bandwidth conditions to a subjective test can influence the nominal scores achieved for the narrowband conditions. The narrowband [ITU-T P.863] scale is provided to give a scale that is compatible with the majority of subjective testing performed in the previous years and objective tools such as PESQ. It ensures that [ITU-T P.863] can be compared to results from narrowband subjective testing.

7.2 When should the narrowband mode be used?

Use the narrowband mode when assessing a narrowband system, analysing results from a narrowband subjective test, or when interested in how quality would be perceived in a traditional telephone network. In addition, use the narrowband mode if comparing [b-ITU-T P.862.1] results with [ITU-T P.863] results.

7.3 When should the superwideband mode be used?

The superwideband mode should be used in all other cases (not covered by clause 7.2). Superwideband mode is also appropriate for assessing narrowband connections. When assessing narrowband connections in superwideband mode, a superwideband reference file is also required as an input to [ITU-T P.863]. [ITU-T P.863] is then able to determine the narrowing of the signal's bandwidth and includes this information in its prediction of quality.

7.4 Why no wideband (up to 7 kHz) mode?

The superwideband scale should be used when assessing a wideband system; the model will account for the 7 kHz bandwidth in its quality assessment. A wideband scale was not provided because the long-term intention is that all quality scores are reported on the superwideband scale. It is intended that [ITU-T P.863] superwideband mode represents a single voice quality scale for all systems regardless of bandwidth limitation and that bandwidth limitation is treated as simply a form of degradation. The narrowband mode has been provided for backward compatibility for [b-ITU-T P.862.1], but backward compatibility was felt to be less important for [b-ITU-T P.862.2] because far less testing is currently performed in wideband than narrowband.

7.5 Are narrowband signals scored equally in narrowband and superwideband modes?

No. Although it is possible to apply an external IRS filter to the signals before passing to [ITU-T P.863] in superwideband mode, the same score is not expected from the superwideband and narrowband modes because of different optimizations used in the perceptual modelling. A narrowband signal with the maximum narrowband mode score of 4.5 will, by design, score around 3.8 in superwideband mode.

7.6 Can I map a narrowband score to a superwideband score?

No. The superwideband scale is not a simple extension of the narrowband or wideband scales to 4.75. It is a completely new scale on which nominal scores for narrowband and wideband codecs are different. A study performed as part of the [ITU-T P.863] characterization phase found that [ITU-T P.863] in superwideband mode will produce relatively lower scores for narrowband and wideband codecs compared to the scores observed in narrowband and wideband subjective tests.

The maximum superwideband scale score of 4.75 was chosen to indicate that superwideband connections offer potentially higher quality than narrowband connections. The narrowband scale produces a maximum score of 4.5.

7.7 Is it recommended to mix bandwidths in subjective testing?

Generally, it is not recommended to mix different bandwidths in a subjective test. When different bandwidth signals are mixed in a single test the bandwidth differences can dominate the resulting mean opinion scores. The [ITU-T P.863] superwideband scale has been engineered to enable easy comparison of signals with different bandwidths which may be difficult to achieve in subjective tests.

8 Influence of reference speech on scores

The quality and content of the reference speech material has an impact on the robustness of the speech quality measurement process. The [ITU-T P.863] model transforms the reference signal to an idealized form; this process is sensitive to voice timbre and background noise.

8.1 What characteristics should a reference signal contain?

Different conditions exist for a superwideband reference signal and a narrowband reference signal. The list below describes a common set of required characteristics for both superwideband and narrowband reference test signals:

- at least three seconds of active speech;
- at least onesecond of silence between active speech periods;
- no more than sixseconds of active speech;
- total length of test sample, including silence, should be no more than 12 seconds;
- active speech level of –26 dBov;
- 16-bit linear pulse code modulation (PCM) encoded;
- noise floor < –80 dBov (A).

Additional characteristics for [ITU-T P.863] superwideband reference test signals are:

- 48 k sample rate;
- filtered 50 Hz to 14 kHz.

Additional characteristics for [ITU-T P.863] narrowband reference test signals are:

- 8 k sample rate;
- filtered 50 Hz to 3.8 kHz.

8.2 Are there constraints on the recording environment?

The room where reference material is recorded must have a reverberation time below 300 ms above 200 Hz (e.g., an anechoic chamber). Recordings must be made using omni-directional microphones. The distance to the microphone must be approximately 10 cm. Background noise must be below 30 dB SPL(A). Directional microphones are allowed on the condition that the spectral balance frequency response is the same as with the omni-directional microphones.

8.3 Can we use artificial speech signals?

An artificial voice signal, such as that defined in [b-ITU-T P.50], is not recommended. Speech signals generated using a text-to-speech system are better, but have had limited validation.

8.4 What filter specification should be used?

A reference signal should be filtered before presenting it to the [ITU-T P.863] model. A different filter is required for the super-wideband and narrowband modes. The filter definitions are provided in Tables 2 and 3.

The superwideband filter definition is approximately described in Table 2; the specific filter definition can be found in [b-ITU-T G.191].

Table 2 – Superwideband filter definition

Frequency (Hz)	Superwideband Gain (dB)
20	–90
50	–3
53	0
13 500	0
14 000	–3
24 000	–90

The approximate narrowband filter definition is given in Table 3.

Table 3 – Narrowband filter definition

Frequency (Hz)	Narrowband Gain (dB)
20	–90
100	–3
103	0
3 750	0
3 800	–3
4 000	–90

8.5 Does reference material influence scores?

Reference speech material has a direct influence on the resulting score. This is true in subjective testing as well as objective testing. A study of 400 speech recordings, from four different speakers, processed through various codec simulations, was found to produce a [ITU-T P.863] superwideband (SWB) condition score range of 1.2 to 1.4 MOS. It is therefore important to use a variety of material when testing a condition's speech quality.

8.6 How much reference material should I use?

[ITU-T P.863] recommends two samples from each of two male and two female speakers, i.e., eight sentence pairs. Some applications may only permit shorter test durations. Mixing of speakers or genders in a single test sample may limit the maximum score that can be achieved.

Assessing multiple speakers and sentence pairs is required to remove material-specific result bias. It has been found that for some low bit-rate codecs the score can differ by as much as 1.4 depending on the selected talker and sentence pair. This material dependency is only removed by ensuring that the scores, of multiple talker and sentence pairs, are averaged.

The factors in the reference material that influence the scores include, for example, talker gender, talker language and signal duration.

8.7 What is the impact of silence padding?

Silence padding is the addition of silence to the front or back of the reference signal. Typically, sentence material in subjective tests has a 0.5 second silence lead in, two sentences, and then a 0.5 second silence at the end of the signal. If these lead-in and end silences are changed, a slight change in score (approximately 0.1) may be seen when comparing the reference signal with itself. This change indicates that the [ITU-T P.863] idealization processing is sensitive for this reference signal. The slight change in score is not a significant problem, but if this effect is seen, another reference recording may be selected to increase absolute consistency in the results.

8.8 How do I validate a reference signal?

One of the main differences between [b-ITU-T P.862] and [ITU-T P.863] is that the [ITU-T P.863] algorithm takes into account signal level, potential acoustic reverberations and bandwidth limiting. This means that the reference signals used with [ITU-T P.863] should not be corrupted by large level variations, a presence of reverberation, or an imbalanced timbre. A simple way to check this is by comparing a candidate reference recording with itself. If the result is not equal to the maximum theoretical score of the model, then the tested recording is not suitable for use as a reference signal for [ITU-T P.863]. In narrowband mode, a reference-to-reference comparison should score 4.5. In superwideband mode, the reference-to-reference comparison should score 4.75.

To further confirm the appropriateness of a reference signal, tests should be performed after adding small offset changes (10 ms and 15 ms) to the start of the clean reference before presenting it to [ITU-T P.863]. The predicted score should remain equal to the maximum theoretical score of the model. A full list of tests required to validate a reference recording is given in Table 4 below.

Table 4 – List of tests required to validate a reference recording

Test ID	Description
1	ITU-T P.56 active speech level = –26 dBov (±2)
2	Noise floor ≤ –80 dBov
3	Total length, including silences ≤ 12 s
4	Minimum duration of active speech ≥ 3
5	Maximum duration of active speech ≤ 6
6	Number of speech segments ≥ 2
7	Silence at start ≥ 0.25 s
8	Silence at end ≥ 0.25 s
9	Longest silence between speech events ≥ 1.0 s
10	Bandpass upper frequency cut-off ≥ 14 000 Hz
11	Superwideband score
12	Superwideband score with 10 ms offset
13	Superwideband score with 15 ms offset
14	Narrowband score
15	Narrowband score with 10 ms offset
16	Narrowband score with 15 ms offset

A validation of the speech material in [b-ITU-T P.501] using these test criteria is presented in Appendix I.

8.9 How do I use a 10 minute reference file?

If one wishes to use a reference speech sample longer than the recommended maximum 6 seconds of active speech, it is recommended that the signal is split into multiple 3 to 6 second active speech sections, and a separate score be computed for each section. Average the scores to determine a single score for the complete reference signal.

9 How to inject test signals

9.1 Over which interfaces can signals be injected into a network?

It is possible to inject a test signal into a network via an acoustic, electrical or digital connection. It is equally possible to inject the test signals digitally into software emulations. Regardless of how the signal is injected into the system under test, filtering of the test signal prior to injection and the level of the signal within the system need to be carefully managed.

9.2 Should the test signal be filtered?

Yes. The signal must be filtered for direct input into the system under test. For example, it could be appropriate to apply an IRS ([b-ITU-T P.48]) filter to a signal applied to a public switched telephone network (PSTN) connection, or to apply a modified IRS (MIRS) filter to a signal applied to a mobile network connection. A signal applied to a wideband network should be low-pass filtered to 7.8 kHz. A signal applied to a superwideband connection may not require any filtering.

Mixed bandwidths might be required for some applications such as conference bridge testing. It may be appropriate to apply IRS, MIRS, wideband and superwideband signals to the different connections into the bridge.

Speech signals applied acoustically via an artificial mouth should not be filtered.

9.3 What level should be used?

The mean active speech level of the signal must be adjusted to the requirements for the system under test.

For example, an electrical signal may be presented at –48 dBm at a microphone input of a voice over IP (VoIP) phone or mobile phone, or it could be at –20 dBm for a PSTN line. The level adjustment must be made after any sample rate changes and filtering are applied.

The effect of different speech levels upon the system under test can be evaluated in superwideband mode. Testing at lower speech levels can be expected to yield lower scores.

10 Influence of recording process on scores

How tests are performed, and in particular how the recorded speech is stored in a file, can influence [ITU-T P.863] results. This clause discusses various aspects that should be controlled when capturing a signal in a test.

10.1 What influence does file speech level have on score?

[ITU-T P.863] was designed to identify the effect of speech level upon speech quality, thus the score will be lower if the speech level is lower when using the superwideband mode.

The score predicted by [ITU-T P.863] in the superwideband mode will be almost unaffected in the range –20 dBov to –36 dBov. The score reduces by approximately 1 MOS between –36 dBov and –46 dBov. [ITU-T P.863] will be less consistent outside this range and is not recommended for signals below –56 dBov.

The score predicted by [ITU-T P.863] in the narrowband mode will be consistent between samples if the signal level is in the range –26 dBov to –46 dBov.

If the mean active speech level of the signal in the degraded file is very low or very high then the score may be affected by the noise floor or amplitude clipping.

10.2 Should we pre-align signal levels?

Level pre-alignment, a process not defined in [ITU-T P.863], describes how a recorded signal's active speech level is normalized to a value of –26 dBov prior to submission to the [ITU-T P.863] algorithm. When applied, level pre-alignment removes the influence of listening level on score prediction.

[ITU-T P.863] assumes that a digital file level of –26 dBov equates to an acoustic listening level of 79 dB SPL in narrowband mode and 73 dB SPL in superwideband mode. The degraded file obtained from a test will rarely have a defined relationship to these sound pressure levels. It is therefore desirable to adjust the level of the recording to –26 dBov so that the influence of speech level on score is eliminated.

If the objective of the test is to evaluate the performance of the system under test at different speech levels, then level pre-alignment should not be applied and the superwideband mode should be used.

Level pre-alignment should always be applied in the [ITU-T P.863] narrowband mode, as the score will be affected and the results will not be predictable. The [ITU-T P.863] narrowband mode validation assumes that all file levels are at –26 dBov. In addition, [b-ITU-T P.862] automatically applied level pre-alignment, thus level pre-alignment should be applied in order to maintain backward compatibility with [b-ITU-T P.862].

10.3 Should we apply a filter to the recording?

[ITU-T P.863] in superwideband mode applies an internal filter that is equivalent to a diffuse field Hi-Fi headset, while in narrowband mode an IRS receive filter is applied. As such, it is not necessary to further filter the degraded file before processing it through [ITU-T P.863]. It is important to manage level and spectral response in the conversion of an analogue signal collected from the network interface for [ITU-T P.863]. The circuits should not filter the signal below 14 kHz or impose any distortion or noise.

10.4 Does ITU-T P.863 handle silence padding?

When recording test signals on a test network, it is common to capture additional silence before and after the test signal. This occurs, for example, when testing between two unsynchronized locations and using large capture windows to compensate for the lack of synchronization. These additional silences at the start and end of the recorded signal should have no influence on the [ITU-T P.863] result. This is the case in narrowband mode, but appears to not be the case in superwideband mode where the score may be affected by some speech material. It is therefore recommended to remove padding from the recording files when presenting these files to the [ITU-T P.863] superwideband model.

10.5 Can I assess a narrowband recording in superwideband mode?

[ITU-T P.863] will evaluate a narrowband degraded speech file in superwideband mode and produce a score prediction that takes into account the lower bandwidth. The maximum score will be around 3.8 on the superwideband scale. The [ITU-T P.863] superwideband mode was designed to enable the evaluation of narrowband, wideband and superwideband signals on the same scale. It relies upon detecting the absence of any speech energy above 3.8 kHz to indicate a narrowband degraded file and the absence of any speech energy above 7 kHz to indicate a wideband degraded file.

11 Behaviour of ITU-T P.863

This clause describes how [ITU-T P.863] scores are affected by various network degrading processes. It is intended that this clause will help in the analysis of [ITU-T P.863] results by providing reference information on how specific degradations influence the final score. It should be noted that in actual network testing it is likely that multiple degrading processes will be responsible for a drop in score and the interaction between these combined degradations may not be linear.

11.1 Global level changes

This clause assumes that level pre-alignment described in clause 10.2 is not applied.

[ITU-T P.863] considers the presentation level of the speech signals in its prediction. Within the specification, and in the evaluation phase, a level range of +5 dB to –20 dB relative to the nominal level was specified and tested.

Figure 1 illustrates the dependency on the presentation level for five different superwideband samples. The average for all five samples is given as a solid bold line.

The nominal level is marked by a vertical dashed line at –26 dBov that corresponds to 73 dB SPL at each ear in a diotic presentation. The white area is the specified range; the grey shaded area is the level range outside of the specification for [ITU-T P.863].

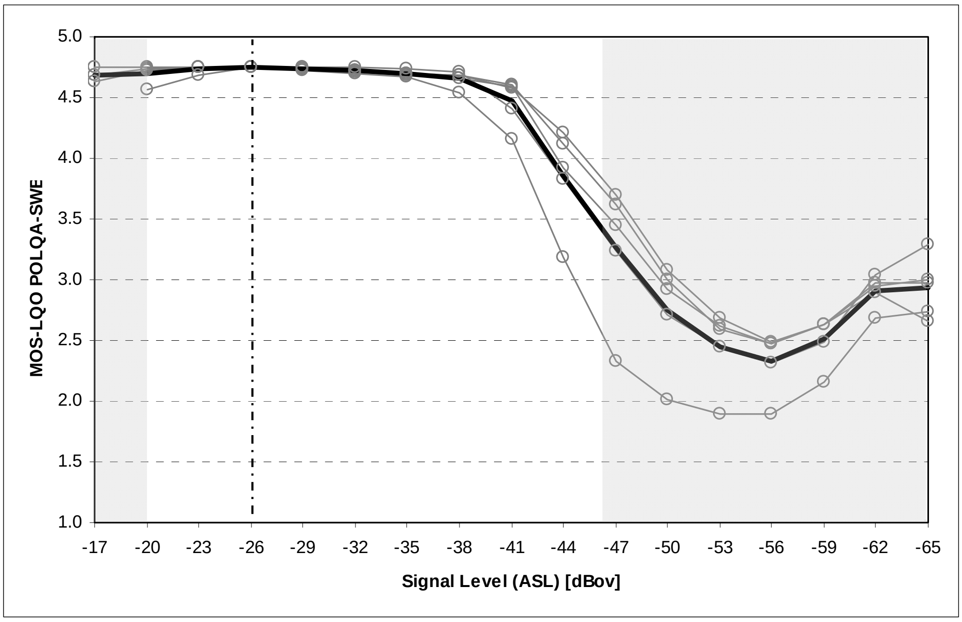

Figure 1 – Dependency of the MOS-LQO$_{SWB}$ on the presentation level

It is recommended that the signal levels in superwideband mode of [ITU-T P.863] neither exceed –20 dBov nor fall below –46 dBov. [ITU-T P.863] must not be used to predict signals of an active speech level (ASL) below –56 dBov.

Presentation level was not required for [ITU-T P.863] narrowband mode and all databases were normalized to –26 dBov. However, for completeness an equivalent graph is shown in Figure 2 for the influence of presentation level in narrowband mode.

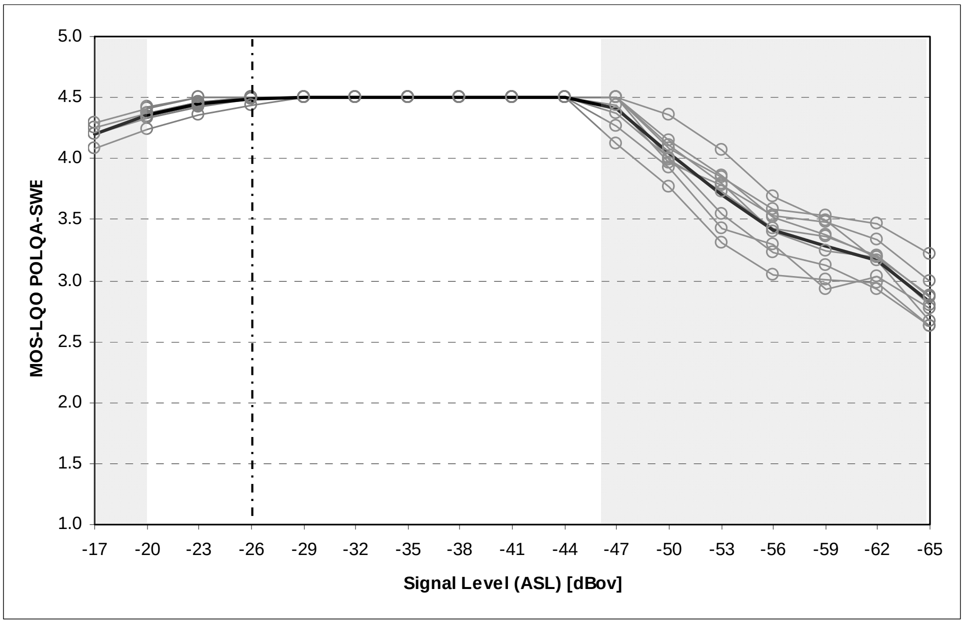

Figure 2 – Influence of the presentation level on MOS-LQO$_{NB}$

It can be seen from the graph that [ITU-T P.863] narrowband mode exhibits almost no level dependency in the range –26 to –46 dBov.

11.2 Local level changes

Discussions of local level changes (automatic gain control (AGC)) are for further study.

11.3 Bandwidth limitations

The score in superwideband mode is affected by bandwidth differences between a degraded signal and its superwideband reference signal. Degraded signals that have some spectral loss between 8 kHz and 14 kHz will have scores below the scale's 4.75 maximum. The superwideband scale will produce a score of approximately 3.8 for a narrowband signal with no other impairment and yield a score around 4.5 for a wideband signal.

It should be noted that [ITU-T P.863] scores may be influenced by the spectral content of the talker's speech in the reference file. If the spectral content is unbalanced, the maximum score for the scale will not be reached.

11.4 Stretching and compressing speech

Temporal compression or stretching in the context of speech coding and voice quality measurement is synonymous with "time scaling". From the perspective of [ITU-T P.863], this can be seen as four distinct categories:

1) Uniform time scaling without pitch preservation.
2) Uniform time scaling with pitch preservation.
3) Non-uniform time scaling without pitch preservation.
4) Non-uniform time scaling with pitch preservation.

A simple modelling of uniform time scaling without pitch preservation can be performed by varying the actual sample rate of a signal while keeping its nominal sample rate constant (i.e., resample a 48 kHz to 48.005 kHz signal and claim that it still has a 48 kHz sample rate). Non-uniform time scaling, is where small sections of the signal are stretched or compressed, or multiple compression and stretching factors are applied to different sections of the speech signal.

Time scaling is frequently observed in modern applications, but may also be caused by the test equipment itself. The sources for time scaling are manifold, but the most frequent ones are:

- Advanced jitter buffer adaptation algorithms in VoIP systems.
- Packet loss/error concealment methods.
- Poor clock generators in A/D or D/A converters.

11.4.1 Uniform time scaling without pitch preservation

A consequence of uniform time scaling is that [ITU-T P.863] sees a continuously increasing or decreasing delay between the reference and the degraded signal. In [ITU-T P.863], this is overcome by an algorithm which estimates the amount of time scaling and subsequently resamples the input signals in order to eliminate the temporal compression or stretching.

[ITU-T P.863] exhibits less than 0.15 MOS deviation for ±3% difference in uniform time scaling between reference and degraded speech files.

11.4.2 Uniform time scaling with pitch preservation

Discussions of uniform time scaling with pitch preservation are for further study.

11.4.3 Non-uniform time scaling without pitch preservation

This type of degradation is unlikely in live networks. As such no studies investigating this topic have been performed.

11.4.4 Non-uniform time scaling with pitch preservation

Studies into this area indicate that [ITU-T P.863] has more difficulty in predicting scores where high levels of non-uniform compression occur, than when non-uniform stretching occurs. With stretching, the scores typically appear to be within 0.5 of the subject test scores, while with compression the scores appear to be as much as 2.0 adrift. In all cases [ITU-T P.863] has a tendency to underestimate the MOS returned from a subjective test.

11.5 Step delay changes

Step changes in delay between reference and degraded speech files result from adaptive jitter buffer implementations which adjust the silences between talk spurts. Such adjustments have a negligible effect in subjective test scores. [ITU-T P.863] will generally detect and account for step changes, but may fail to cope with detecting and eliminating all time-shifting, and thus return low scores. The ability to compensate for time-shifting depends on the magnitude of the time-shift and also the location within the sample. Results also seem to differ between male and female samples.

11.6 Long-term gaps in speech

[ITU-T P.863] identifies missing speech, whether at the beginning, the middle or the end of a sentence. The score will be reduced whenever speech is missing.

Periods of silence might be in the range 200-500 ms. The cause of such silences might be an inter-system handover (from 3G to 2G) in mobile connections or sustained packet loss due to re-buffering in VoIP or streaming applications. The subjective effect will depend on whether the silence occurs at the beginning, the middle or the end of a sentence.

Loss of speech in the middle of a sentence will affect the score more than loss at the beginning or the end of a sentence.

11.7 Short-term gaps in speech

[ITU-T P.863] considers the amount of lost speech in its MOS prediction. The subjective perception of lost speech is highly dependent on the occurrence of the loss with regard to the speech information content. It is dependent on the structure of the loss pattern in conjunction with the temporal structure of the particular speech file.

Short-term loss of speech might be caused by packet loss giving rise to occasional losses of 20 ms or more.

[ITU-T P.863] indicates a reduction in score for small loss rates. There is some dependency on different speech content of the reference material. Loss rates of 25% and above result in a range close to a score of 1, which was a requirement of the specification for [ITU-T P.863].

11.8 Cellular handovers

Discussions of cellular handovers are for further study.

12 Comparison of ITU-T P.862.1/ITU-T P.862.2 and ITU-T P.863

12.1 Scale differences

This clause provides guidelines on how to compare existing speech quality results obtained with [b-ITU-T P.862.1] and [ITU-T P.863] narrowband mode, as well as how to compare existing results from [b-ITU-T P.862.2] and [ITU-T P.863] superwideband mode.

[ITU-T P.863] will return scores very similar to PESQ [b-ITU-T P.862.1] in the narrowband mode with simple codecs such as ITU-T G.711. Tests with more sophisticated codecs and transmission techniques may yield different scores as [ITU-T P.863] addresses the objective assessment limitations of PESQ.

It is more difficult to compare [ITU-T P.863] superwideband mode with PESQ [b-ITU-T P.862.2] because most wideband experiments were performed with 16 k sample rate material. [ITU-T P.863] superwideband mode requires a 48 k sample rate reference file. The [ITU-T P.863] results with a 16 k sample rate reference file or an up-sampled 16 k reference file will be wrong, as there will not be any speech energy above 8 kHz.

12.2 Degradation handling differences

Table 5 provides a brief explanation of what can be expected from [ITU-T P.863] for certain types of degradation compared to the PESQ processing model.

Table 5 – Brief explanation

Degradation	[ITU-T P.863]	[b-ITU-T P.862]/[b-ITU-T P.862.1]
Bandwidth limitations	[ITU-T P.863] in superwideband mode takes into account bandwidth limitations by detecting the absence of any speech energy above 3.8 kHz to indicate a narrowband degraded file and the absence of any speech energy above 7 kHz to indicate a wideband degraded file. With a narrowband signal the maximum achievable score is 3.8. Change in bandwidth between speakers (different gender or talker) in a single test may lead to a lower than expected score.	PESQ applies a linear frequency equalization stage before presenting the signals to the psycho-acoustic model. This effectively removes frequency response influences from being detected in the model. This is useful for small degrees of frequency shaping, but PESQ underestimates severe linear frequency response distortions.

Table 5 – Brief explanation

Degradation	[ITU-T P.863]	[b-ITU-T P.862]/[b-ITU-T P.862.1]
Short interrupts (e.g., packet loss)	The predicted quality score tends to a MOS value of 1.0 as frame loss increases up to 30%. Even small loss rates cause a drop in measured speech quality. Results show that the superwideband mode is slightly more sensitive to short interrupts than the narrowband mode.	PESQ behaves in a similar way to [ITU-T P.863] with scores tending to a MOS of 1.0 as loss rate increases to 30%.
Long interrupts (e.g., VAD clipping, inter-system handovers)	Long interrupts describe muting of speech for 200 ms or more at the front, in the middle or the end of a speech sentence. Loss in the middle of speech leads to the largest drop in quality, followed by front-end loss with losses at the end of a sentence having least impact on score.	It has been claimed that PESQ reacts unexpectedly to lost speech. For small interrupts [ITU-T P.863] and PESQ produce consistent scores, but with longer interrupts PESQ predictions are significantly more optimistic than expected.

12.3 Standard codecs

Table 6 presents average scores for a number of narrowband codecs obtained from the data presented in the [b-ITU-T P.862] application guide as well as average scores, with pre-alignment applied, taken from Appendix II. Please note that the scores are only presented to one decimal place because users are unable to distinguish differences of less than one decimal place.

Table 6 – Average scores

Codec	[b-ITU-T P.862.1]	[ITU-T P.863] NB	[ITU-T P.863] SWB
ITU-T G.711 a-law	4.4	4.4	3.4
ITU-T G.711 µ-law	4.5	4.4	3.4
ITU-T G.726 32 kbit/s	4.1	4.4	3.2
ITU-T G.728	4.1	4.3	3.0
ITU-T G.729	3.9	4.2	3.0
ITU-T G.729A	3.8	4.1	2.9
AMR 12.2 kbit/s	4.0	4.4	3.2
AMR 10.2 kbit/s	3.9	4.3	3.1
AMR 7.95 kbit/s	3.7	4.1	2.9
AMR 7.4 kbit/s	3.7	4.1	2.9
AMR 6.7 kbit/s	3.6	4.0	2.9
AMR 5.9 kbit/s	3.5	3.8	2.7

From Table 6 it can be seen that the [ITU-T P.863] narrowband (NB) scale compresses the upper part of the narrowband scale. In [b-ITU-T P.862], codec quality ranged from 4.5 to 3.5 while in [ITU-T P.863] the range is reduced to between 4.4 and 3.8. The rank order is not changed. Although the compression of the upper part of the scale is seen, the rank order is generally maintained. The only change in rank order occurs between the AMR 12.2 kbit/s and G.728 and even this change is insignificant.

The scores obtained for these narrowband codecs are around 0.9 lower on the [ITU-T P.863] SWB scale than on the [b-ITU-T P.862] scale.

13 Procedure for comparing subjective test results to ITU-T P.863 results

It has never been possible to directly compare the results of two or more subjective tests because too many variables influence the results for a specific condition. Even if the same subjective test is run in the same subjective test laboratory with two different groups of people, slight differences in condition scores are always expected. The reason is that one group will never vote exactly the same way as a second group. When further variables are added, such as a different condition order, or different types of degradation, the absolute score achieved for a given condition will change.

Naturally, from a business perspective it is important to be able to compare the results of a new test with results from previous tests, and subjective test experts have developed various mechanisms to support this interpretation. These include the use of common reference conditions between tests, and also careful balancing of the test design to ensure a reasonable range of quality. Despite this effort, different subjective tests will always exhibit a broad range of scores for a given condition.

The very nature of an objective tool is that it always reports the same absolute score for a given speech recording. This is opposite to how a subjective test behaves. Therefore, a process must be used to compensate for the methodology mismatch between a subjective and an objective result set.

Before comparing results of a subjective test with [ITU-T P.863] results, perform the following simple linear-normalization process:

1) Calculate the linear-regression function between subjective test condition average scores (Subj) and [ITU-T P.863] condition average scores (P863) using the formula below:

   a. $\textit{regression formula(y)} = a + bx$

   b. $\textit{slope(b)} = \left[N * \sum(Subj * P863) - \sum(Subj) * \sum(P863)\right] / \left[N * \sum\left(Subj^2\right) - \left(\sum Subj\right)^2\right]$

   c. $\textit{intercept(a)} = \left[\sum(P863) - b * \sum(Subj)\right] / N$

2) Apply the linear-regression formula calculated above to each subjective test condition average.

Table 7 shows a worked example:

Table 7 – Worked example for mapping objective to subjective scores

Condition	Subj	ITU-T P863	Subj2	Subj*ITU-T P863	Mapped-Subj
1	3.52	3.95	12.40	13.89	4.05
2	2.84	3.55	8.09	10.09	3.35
3	2.30	3.02	5.30	6.96	2.80
4	3.39	4.20	11.46	14.22	3.91
5	2.69	3.19	7.22	8.56	3.19
6	3.19	3.79	10.16	12.08	3.70
7	3.56	4.45	12.69	15.87	4.09
8	3.28	3.53	10.77	11.60	3.80
9	3.27	3.51	10.70	11.49	3.79
10	2.69	3.03	7.22	8.14	3.19
11	3.17	3.64	10.03	11.51	3.68
12	2.91	3.43	8.45	9.98	3.42
13	2.64	3.08	6.95	8.11	3.14
14	2.94	3.01	8.63	8.84	3.45
15	2.24	2.88	5.02	6.46	2.73

Table 7 – Worked example for mapping objective to subjective scores

Condition	Subj	ITU-T P863	Subj2	Subj*ITU-T P863	Mapped-Subj
16	3.29	3.75	10.84	12.34	3.81
17	3.19	3.57	10.16	11.37	3.70
18	2.63	3.34	6.89	8.78	3.13
19	2.78	3.10	7.74	8.61	3.29
20	2.28	2.75	5.20	6.27	2.78
21	2.28	2.83	5.20	6.46	2.78
22	2.38	2.50	5.64	5.95	2.87
23	1.79	2.46	3.21	4.41	2.27
24	2.52	2.96	6.35	7.45	3.02
25	2.32	2.69	5.40	6.25	2.82
26	2.33	3.16	5.44	7.38	2.83
27	2.33	2.69	5.44	6.28	2.83
28	1.75	2.42	3.06	4.23	2.23
29	2.75	3.36	7.56	9.25	3.26
30	2.63	3.07	6.89	8.05	3.13
31	2.53	2.91	6.41	7.37	3.03
32	2.32	2.56	5.40	5.94	2.82
33	2.07	2.51	4.30	5.20	2.56
34	2.60	2.77	6.78	7.22	3.11
35	1.70	2.28	2.88	3.87	2.18
36	1.73	2.37	2.99	4.10	2.21
37	1.10	1.50	1.22	1.65	1.57
38	1.42	1.95	2.01	2.76	1.89
39	2.24	2.82	5.02	6.31	2.73
40	3.02	3.52	9.13	10.65	3.53
41	3.51	4.12	12.32	14.47	4.03
42	3.72	4.38	13.83	16.29	4.25
43	3.68	4.42	13.52	16.24	4.21
44	3.80	4.45	14.46	16.93	4.33
sum	117.31	139.46	330.36	389.85	

slope (b) = (44 * 389.85 – 117.31 * 139.46)/(44 * 330.36 – 117.31^2) = 1.0247

intercept(a) = (139.46 – 1.0247 * 117.31)/44 = 0.4376

In Figure 3, the graph on the left plots un-normalized subjective test scores against [ITU-T P.863], while the graph on the right illustrates the effect of normalizing/mapping the subjective scores to [ITU-T P.863].

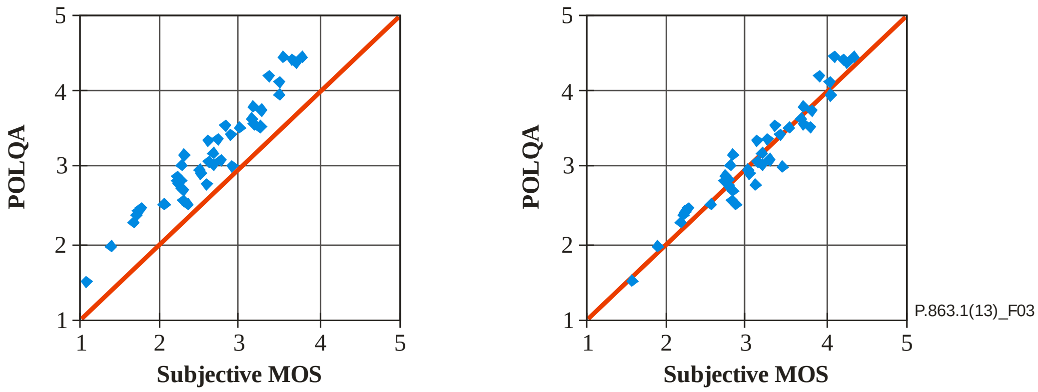

Figure 3 – Illustration of mapping

The use of linear regression removes simple test condition biases from the subjective results.

Further normalization is possible by applying third-order monotonic polynomial regression function to remove biases that occur when a subjective test does not have a broad enough spread of quality, but applying this linear regression normalization process is a minimum requirement.

The topic of comparing subjective test results with objective model predictions is more fully discussed in [b-ITU-T P.1401].

14 Validation scope

[ITU-T P.863] has been validated across many degradation types that are seen today on telephone networks, or expected to become prevalent over the next 10 years.

14.1 Validated

Table 8 provides a list of factors that have been used in the selection and validation phase of the [ITU-T P.863] algorithm.

Table 8 – Validated factors

Test factors
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment
Bit rates if a codec has more than one bit-rate mode
Transcoding
Acoustic noise in sending environment
Effect of varying delay in listening only tests
Short-term time warping of audio signal
Long-term time warping of audio signal
Listening levels between 53 and 78 dB(A) SPL in superwideband mode
Packet loss and packet loss concealment with PCM type codecs
Temporal and amplitude clipping of speech
Linear distortions, including bandwidth limitations and spectral shaping ('non-flat frequency responses')
Frequency response

Table 8 – Validated factors

Coding technologies
ITU-T G.711, ITU-T G.711 PLC, ITU-T G.711.1
ITU-T G.718, ITU-T G.719, ITU-T G.722, ITU-T G.722.1, ITU-T G.723.1, ITU-T G.726, ITU-T G.728, ITU-T G.729
GSM-FR, GSM-HR, GSM EFR
AMR-NB, AMR-WB (ITU-T G.722.2), AMR-WB+
PDC-FR, PDC-HR
EVRC (ANSI/TIA-127-A), EVRC-B (TIA-718-B)
Skype (SILK V3, iLBC, iSAC and ITU-T G.729)
Speex, QCELP (TIA-EIA-IS-733), iLBC, CVSD (64 kbit/s, "Bluetooth")
MP3, AAC, AAC-LD
Applications
Test Factors
Codec evaluation
Terminal testing, influence of the acoustic path and the transducer in the sending and receiving direction. (NOTE – Acoustic path in receiving direction only for SWB mode)
Bandwidth extensions
Live network testing using digital or analogue connection to the network
Testing of emulated and prototype networks
UMTS, CDMA, GSM, TETRA, WB-DECT, VoIP, POTS, PSTN, Video Telephony, Bluetooth
Voice Activity Detection (VAD), Automatic Gain Control (AGC)
Voice Enhancement Devices (VED), Noise Reduction (NR)
Discontinuous Transmission (DTX), Comfort Noise Insertion

14.2 Not yet validated

[ITU-T P.863] has not been validated against the variables given in Table 9.

Table 9 – Factors not validated

Test factors
Talker dependencies
Multiple simultaneous talkers
Bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode
Network information signals as input to a codec
Artificial speech signals as input to a codec
Music as input to a codec
Listener echo
Coding Technologies
Coding technologies operating below 4 kbit/s

14.3 Outside scope

[ITU-T P.863] is not intended to be used with the variables provided in Table 10.

Table 10 – Factors outside scope

Test factors
Effect of delay in conversational tests
Talker echo
Sidetone
Acoustic noise in receiving environment
Applications
Non-intrusive measurements
Two-way communications performance

Appendix I

ITU-T P.863 Reference validation checks

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

Table I.1 contains [ITU-T P.863] reference validation checks for the speech samples contained in [b-ITU-T P.501]. These files are indicated in [ITU-T P.863] as possible reference files for [ITU-T P.863] testing. Unfortunately, most files in [b-ITU-T P.501] fail to meet the reference requirements specified in clause 8.1. It may be possible to fix the failing files by pre-processing them with, for example, level normalization, de-essing and global frequency normalization.

Table I.1 – Reference validation checks

Language	File	Valid	Failing tests
American English	female 1	No	1, 2, 4, 7, 9-13
	female 2	No	1, 2, 4, 7, 9-13
	male 1	No	1, 2, 4, 7, 9-13
	male 2	No	1, 2, 4, 8-13
Chinese	female 1	No	11
	female 2	No	1, 11-13
	male 1	No	11, 13
	male 2	No	1
Dutch FB	female 1	No	12
	female 2	No	12
	male 1	Yes	
	male 2	Yes	
English	female 1	Yes	
	female 2	No	1, 13
	male 1	No	11-13
	male 2	Yes	
Finnish	female 1	No	1, 2, 5, 14-16
	female 2	No	1, 2, 5, 11-16
	male 1	No	1, 2, 5, 9, 11, 14-16
	male 2	No	1, 2, 12, 14-16
French	female 1	No	1, 11-13
	female 2	No	2, 11-13
	male 1	No	1, 3, 5, 11-14, 16
	male 2	No	2, 11-14
German FB	female 1	No	2, 9
	female 2	No	2, 9
	male 1	No	2, 4, 7, 9
	male 2	No	2, 11, 13

Table I.1 – Reference validation checks

Language	File	Valid	Failing tests
German	female 1	No	1, 2, 4, 6, 7, 9-14
	female 2	No	1, 2, 4, 7, 9-13
	male 1	No	1, 2, 4, 7, 9-13
	male 2	No	1, 2, 4, 7, 9-13
Italian	female 1	No	1, 3, 5, 11, 13-16
	female 2	No	1, 3, 5, 14-16
	male 1	No	1, 3, 5, 11, 12, 14-16
	male 2	No	1, 3, 5, 14-16
Japanese	female 1	No	11
	female 2	Yes	
	male 1	Yes	
	male 2	No	1, 2
Polish	female 1	No	1, 2, 4, 11-13
	female 2	No	2, 4
	male 1	No	1, 2, 4
	male 2	No	2, 4
Spanish (US)	female 1	No	1, 2, 4, 6, 7, 9, 10-13
	female 2	No	1, 2, 4, 6, 10-13
	male 1	No	1, 2, 4, 6, 9-11
	male 2	No	1, 2, 4, 6, 9-11

Appendix II

Typical scores expected from ITU-T P.863 for a given codec

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

Tables II.1 and II.2 list typical scores expected from [ITU-T P.863] for a given codec. It should be noted that because [ITU-T P.863] demonstrates a significant material dependency, as discussed in clause 8, a single sample may differ from the typical value presented below by as much as 1.4 MOS.

Table II.1 – Narrowband mode

Codec name	Details	Without level pre-alignment			With level pre-alignment		
		Minimum	Average	Maximum	Minimum	Average	Maximum
ITU-T G.711	A law, 64 kbit/s	3.98	4.36	4.50	4.00	4.38	4.50
ITU-T G.711	μ law, 64 kbit/s	3.97	4.43	4.50	3.99	4.44	4.50
ITU-T G.726	32 kbit/s	3.90	4.34	4.50	3.93	4.36	4.50
ITU-T G.728	16 kbit/s	3.87	4.34	4.49	3.87	4.34	4.49
ITU-T G.729	8 kbit/s	3.86	4.22	4.50	3.88	4.22	4.49
ITU-T G.729A	8 kbit/s	3.69	4.06	4.44	3.66	4.05	4.43
GSM FR	13 kbit/s	2.97	3.84	4.45	2.94	3.86	4.42
GSM HR	5.6 kbit/s	2.96	3.51	3.85	2.96	3.50	3.88
GSM HR	5.6 kbit/s, DTX on	2.96	3.51	3.85	2.96	3.50	3.88
GSM EFR	12.2 kbit/s	4.11	4.38	4.50	4.12	4.39	4.50
GSM EFR	12.2 kbit/s, DTX on	4.21	4.39	4.50	4.21	4.40	4.50
AMR	4.75 kbit/s	3.04	3.61	4.09	3.03	3.59	4.01
AMR	4.75 kbit/s, DTX on	2.86	3.61	4.09	2.85	3.59	4.07
AMR	5.15 kbit/s	3.00	3.70	4.17	3.01	3.68	4.13
AMR	5.15 kbit/s, DTX on	3.14	3.70	4.15	3.12	3.70	4.16
AMR	5.9 kbit/s	3.32	3.86	4.24	3.30	3.84	4.24
AMR	5.9 kbit/s, DTX on	3.33	3.85	4.27	3.36	3.84	4.25
AMR	6.7 kbit/s	3.59	4.00	4.35	3.57	3.99	4.35
AMR	6.7 kbit/s, DTX on	3.42	3.97	4.38	3.38	3.98	4.35
AMR	7.4 kbit/s	3.69	4.11	4.40	3.83	4.12	4.37
AMR	7.4 kbit/s, DTX on	3.74	4.10	4.46	3.78	4.12	4.44
AMR	7.95 kbit/s	3.82	4.12	4.36	3.79	4.11	4.37
AMR	7.95 kbit/s, DTX on	3.68	4.08	4.39	3.51	4.09	4.40
AMR	10.2 kbit/s	4.06	4.29	4.47	4.07	4.30	4.48
AMR	10.2 kbit/s, DTX on	4.04	4.28	4.50	4.01	4.30	4.50
AMR	12.2 kbit/s	4.10	4.35	4.50	4.12	4.36	4.50
AMR	12.2 kbit/s, DTX on	4.04	4.32	4.50	4.06	4.34	4.50

Table II.2 – Superwideband mode

Codec name	Details	Without level pre-alignment			With level pre-alignment		
		Minimum	Average	Maximum	Minimum	Average	Maximum
ITU-T G.711	A law, 64 kbit/s	2.78	3.34	3.84	2.79	3.34	3.80
ITU-T G.711	µ law, 64 kbit/s	2.72	3.36	3.83	2.74	3.35	3.83
ITU-T G.726	32 kbit/s	2.75	3.23	3.74	2.78	3.23	3.68
ITU-T G.728	16 kbit/s	2.44	3.06	3.65	2.48	3.01	3.51
ITU-T G.729	8 kbit/s	2.61	3.06	3.68	2.65	3.07	3.64
ITU-T G.729A	8 kbit/s	2.33	2.86	3.62	2.37	2.86	3.57
GSM FR	13 kbit/s	1.53	2.44	3.42	1.61	2.50	3.38
GSM HR	5.6 kbit/s	2.05	2.43	2.95	2.09	2.42	2.89
GSM HR	5.6 kbit/s, DTX on	2.05	2.43	2.95	2.09	2.42	2.89
GSM EFR	12.2 kbit/s	2.67	3.20	3.87	2.65	3.22	3.82
GSM EFR	12.2 kbit/s, DTX on	2.64	3.18	3.86	2.63	3.21	3.79
AMR	4.75 kbit/s	1.74	2.48	3.19	1.75	2.47	3.14
AMR	4.75 kbit/s, DTX on	1.75	2.46	3.23	1.78	2.47	3.21
AMR	5.15 kbit/s	1.87	2.55	3.26	1.87	2.55	3.22
AMR	5.15 kbit/s, DTX on	1.83	2.53	3.24	1.84	2.54	3.23
AMR	5.9 kbit/s	2.04	2.69	3.47	2.08	2.70	3.45
AMR	5.9 kbit/s, DTX on	2.01	2.70	3.53	2.01	2.71	3.52
AMR	6.7 kbit/s	2.08	2.84	3.63	2.13	2.86	3.59
AMR	6.7 kbit/s, DTX on	2.11	2.81	3.64	2.13	2.82	3.62
AMR	7.4 kbit/s	2.35	2.94	3.66	2.39	2.95	3.61
AMR	7.4 kbit/s, DTX on	2.26	2.91	3.70	2.28	2.94	3.69
AMR	7.95 kbit/s	2.37	2.91	3.67	2.39	2.92	3.59
AMR	7.95 kbit/s, DTX on	2.34	2.91	3.68	2.46	2.92	3.64
AMR	10.2 kbit/s	2.56	3.12	3.84	2.58	3.13	3.79
AMR	10.2 kbit/s, DTX on	2.57	3.14	3.82	2.57	3.14	3.79
AMR	12.2 kbit/s	2.69	3.19	3.89	2.72	3.21	3.84
AMR	12.2 kbit/s, DTX on	2.70	3.20	3.89	2.73	3.23	3.85
G.711	A law, 64 kbit/s, IRS	3.20	3.47	3.88	2.69	3.30	3.63
G.711	µ law, 64 kbit/s, IRS	2.95	3.38	3.68	2.68	3.29	3.68
G.726	32 kbit/s, IRS	3.09	3.37	3.68	2.67	3.19	3.58
G.728	16 kbit/s, IRS	2.34	2.92	3.44	2.14	2.83	3.42
G.729	8 kbit/s, IRS	2.30	2.93	3.53	2.18	2.86	3.53
G.729	Annex A, 8 kbit/s, IRS	2.16	2.82	3.47	2.05	2.75	3.47
GSM FR	13 kbit/s, IRS	1.60	2.39	2.98	1.49	2.29	2.95
GSM HR	5.6 kbit/s, IRS	1.85	2.45	3.01	1.81	2.39	3.00
GSM HR	5.6 kbit/s, DTX on, IRS	1.85	2.45	3.01	1.81	2.39	3.00
GSM EFR	12.2 kbit/s, IRS	2.57	3.11	3.87	2.65	3.22	3.82
GSM EFR	12.2 kbit/s, DTX on, IRS	2.36	3.06	3.64	2.19	2.98	3.63
AMR	4.75 kbit/s, IRS	1.81	2.49	3.15	1.70	2.41	3.13
AMR	4.75 kbit/s, DTX on, IRS	1.84	2.52	3.08	1.74	2.41	3.04
AMR	5.15 kbit/s, IRS	1.84	2.53	3.14	1.75	2.45	3.17

Table II.2 – Superwideband mode

Codec name	Details	Without level pre-alignment			With level pre-alignment		
		Minimum	Average	Maximum	Minimum	Average	Maximum
AMR	5.15 kbit/s, DTX on, IRS	1.97	2.56	3.14	1.86	2.48	3.14
AMR	5.9 kbit/s, IRS	1.95	2.67	3.38	1.90	2.58	3.37
AMR	5.9 kbit/s, DTX on, IRS	2.00	2.70	3.38	1.87	2.59	3.36
AMR	6.7 kbit/s, IRS	1.91	2.78	3.48	1.83	2.71	3.47
AMR	6.7 kbit/s, DTX on, IRS	1.98	2.77	3.48	1.89	2.68	3.46
AMR	7.4 kbit/s, IRS	2.12	2.84	3.54	2.01	2.77	3.52
AMR	7.4 kbit/s, DTX on, IRS	2.22	2.83	3.48	2.02	2.75	3.51
AMR	7.95 kbit/s, IRS	2.18	2.84	3.57	2.01	2.76	3.55
AMR	7.95 kbit/s, DTX on, IRS	2.19	2.85	3.56	2.03	2.75	3.54
AMR	10.2 kbit/s, IRS	2.33	2.99	3.63	2.13	2.91	3.64
AMR	10.2 kbit/s, DTX on, IRS	2.25	2.98	3.63	2.08	2.88	3.64
AMR	12.2 kbit/s, IRS	2.38	3.08	3.73	2.22	3.00	3.74
AMR	12.2 kbit/s, DTX on, IRS	2.37	3.07	3.70	2.23	2.98	3.69
ITU-T G.722	64 kbit/s	4.01	4.37	4.51	4.04	4.34	4.46
ITU-T G.711.1	A law, 96 kbit/s	4.01	4.37	4.61	4.02	4.36	4.54
ITU-T G.711.1	µ law, 96 kbit/s	4.02	4.38	4.63	4.02	4.37	4.56
ITU-T G.718	32 kbit/s	3.82	4.19	4.49	3.83	4.21	4.46
ITU-T G.718	32 kbit/s, DTX on	3.78	4.20	4.49	3.78	4.21	4.46
ITU-T G.722.1	32 kbit/s	3.13	3.80	4.38	3.07	3.79	4.38
ITU-T G.729.1	32 kbit/s	3.81	4.13	4.49	3.84	4.13	4.46
ITU-T G.722.2	6.6 kbit/s	2.53	3.04	3.54	2.51	3.04	3.50
ITU-T G.722.2	6.6 kbit/s, DTX on	2.52	3.07	3.61	2.52	3.08	3.58
ITU-T G.722.2	8.85 kbit/s	2.81	3.43	3.98	2.84	3.45	4.00
ITU-T G.722.2	8.85 kbit/s, DTX on	2.87	3.47	4.01	2.92	3.49	3.98
ITU-T G.722.2	12.65 kbit/s	3.18	3.73	4.24	3.24	3.77	4.19
ITU-T G.722.2	12.65 kbit/s, DTX on	3.22	3.76	4.26	3.23	3.79	4.21
ITU-T G.722.2	14.25 kbit/s	3.23	3.81	4.28	3.28	3.84	4.26
ITU-T G.722.2	14.25 kbit/s, DTX on	3.30	3.85	4.34	3.26	3.87	4.30
ITU-T G.722.2	15.85 kbit/s	3.28	3.84	4.34	3.32	3.86	4.29
ITU-T G.722.2	15.85 kbit/s, DTX on	3.36	3.90	4.38	3.34	3.92	4.34
ITU-T G.722.2	18.25 kbit/s	3.47	3.96	4.40	3.47	3.97	4.36
ITU-T G.722.2	18.25 kbit/s, DTX on	3.41	3.96	4.43	3.38	3.98	4.41
ITU-T G.722.2	19.85 kbit/s	3.48	3.95	4.41	3.47	3.98	4.37
ITU-T G.722.2	19.85 kbit/s, DTX on	3.43	3.98	4.45	3.41	4.00	4.43
ITU-T G.722.2	23.05 kbit/s	3.55	4.02	4.47	3.58	4.04	4.46
ITU-T G.722.2	23.05 kbit/s, DTX on	3.55	4.05	4.48	3.60	4.06	4.47
ITU-T G.722.2	23.85 kbit/s	3.42	3.97	4.45	3.51	3.99	4.41
ITU-T G.722.2	23.85 kbit/s, DTX on	3.49	4.01	4.49	3.53	4.02	4.47
ITU-T G.718B	48 kbit/s	4.31	4.56	4.75	4.29	4.56	4.75
ITU-T G.718B	48 kbit/s, DTX on	4.35	4.59	4.75	4.31	4.58	4.75
ITU-T G.711.1D	A law, 128 kbit/s	4.57	4.73	4.75	4.53	4.71	4.75

Table II.2 – Superwideband mode

Codec name	Details	Without level pre-alignment			With level pre-alignment		
		Minimum	Average	Maximum	Minimum	Average	Maximum
ITU-T G.711.1D	µ law, 128 kbit/s	4.58	4.73	4.75	4.57	4.72	4.75
ITU-T G.722B	96 kbit/s	4.63	4.74	4.75	4.59	4.72	4.75
ITU-T G.722.1C	48 kbit/s	3.58	4.23	4.74	3.54	4.22	4.73
ITU-T G.729.1E	64 kbit/s	4.56	4.70	4.75	4.53	4.69	4.75

Appendix III

Applications for further investigation

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

Claims have been made that [ITU-T P.863] provides inaccurate predictions of absolute quality for the following:

- Acoustic recordings using free-field microphones without HATS or ear-canal simulation;
- EVRC family codecs (e.g., EVRC, EVRC-B, EVRC-WB, EVRC-NW).

These claims need to be validated by future investigations.

Bibliography

[b-ITU-T G.107] Recommendation ITU-T G.107 (2011), *The E-model: a computational model for use in transmission planning.*

[b-ITU-T G.191] Recommendation ITU-T G.191 (2010), *Software tools for speech and audio coding standardization.*

[b-ITU-T P.48] Recommendation ITU-T P.48 (1988), *Specification for an intermediate reference system.*

[b-ITU-T P.50] Recommendation ITU-T P.50 (1999), *Artificial voices.*

[b-ITU-T P.501] Recommendation ITU-T P.501 (2012), *Test signals for use in telephonometry.*

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

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

[b-ITU-T P.800.1] Recommendation ITU-T P.800.1 (2006), *Mean Opinion Score (MOS) terminology*.

[b-ITU-T P.830] Recommendation ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs.*

[b-ITU-T P.861] Recommendation ITU-T P.861 (1998), *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*. (superseded)

[b-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*.

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

[b-ITU-T P.862.2] Recommendation ITU-T P.862.2 (2007), *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs.*

[b-ITU-T P.1401] Recommendation ITU-T P.1401 (2012), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.*

[b-ITU-T P.Imp863] ITU-T Implementers' Guide II for ITU-T P.863 (2011), *Bug Fixes (zipped with software – conformance test results).*

[b-ITU-R BS.1387] Recommendation ITU-R BS.1387 (2001), *Method for objective measurements of perceived audio quality.*

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Terminals and subjective and objective assessment methods**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2014