

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.863.1

(09/2014)

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
Methods for objective and subjective assessment of quality of services other than voice services	Series	P.1500

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	Unique ID*
1.0	ITU-T P.863.1	2013-05-14	12	11.1002/1000/11935
2.0	ITU-T P.863.1	2014-09-11	12	11.1002/1000/12175

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

FOREWORD

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

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

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

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

NOTE

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

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

INTELLECTUAL PROPERTY RIGHTS

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

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

© ITU 2015

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

Table of Contents

		Page
1	Scope.....	1
2	References.....	1
3	Definitions	1
	3.1 Terms defined elsewhere	1
	3.2 Terms defined in this Recommendation.....	1
4	Abbreviations and acronyms	1
5	Conventions	2
6	Introduction to ITU-T P.863.....	2
	6.1 History of objective speech quality	2
	6.2 Basics of a subjective test.....	3
	6.3 Benefit of using Recommendation ITU-T P.863	4
	6.4 Relationship with other Recommendations.....	4
	6.5 Moving from ITU-T P.862 to ITU-T P.863	4
	6.6 Challenges with ITU-T P.863.....	4
	6.7 MOS misconceptions.....	5
7	Operational modes	5
	7.1 Why two operational modes?	5
	7.2 When should the narrowband mode be used?	6
	7.3 When should the super-wideband mode be used?.....	6
	7.4 Why no wideband (up to 7 kHz) mode?.....	6
	7.5 Are narrowband signals scored equally in narrowband and super- wideband modes?	6
	7.6 Can I map a narrowband score to a super-wideband score?	6
	7.7 Is it recommended to mix bandwidths in subjective testing?.....	6
8	Influence of reference speech on scores	7
	8.1 What characteristics should a reference signal contain?	7
	8.2 Are there constraints on the recording environment?.....	7
	8.3 Can we use artificial speech signals?	7
	8.4 What filter specification should be used?.....	7
	8.5 Does reference material influence scores?	8
	8.6 How much reference material should I use?	8
	8.7 What is the impact of silence padding?	8
	8.8 How do I validate a reference signal?	9
	8.9 How do I use a 10 minute reference file?.....	9
9	How to inject test signals.....	10
	9.1 Over which interfaces can signals be injected into a network?.....	10
	9.2 Should the test signal be filtered?.....	10

9.3	What level should be used?	10
10	Influence of recording process on scores.....	10
10.1	What influence does file speech level have on score?	10
10.2	Should we pre-align signal levels?	11
10.3	Should we apply a filter to the recording?.....	11
10.4	Does silence padding affect the ITU-T P.863 scores?	11
10.5	Can I assess a narrowband recording in super-wideband mode?	11
11	Behaviour of ITU-T P.863.....	12
11.1	Global level changes.....	12
11.2	Local level changes	13
11.3	Bandwidth limitations	13
11.4	Stretching and compressing speech.....	13
11.5	Step delay changes	14
11.6	Long-term gaps in speech.....	14
11.7	Short-term gaps in speech	15
11.8	Cellular handovers.....	15
12	Comparison of ITU-T P.862.1/ITU-T P.862.2 and ITU-T P.863.....	15
12.1	Scale differences.....	15
12.2	Degradation handling differences.....	15
12.3	Standard codecs	16
13	Procedure for comparing subjective test results to ITU-T P.863 results	17
14	Validation scope	19
14.1	Validated.....	19
14.2	Not yet validated.....	20
14.3	Outside scope.....	21
	Appendix I – ITU-T P.863 Transparency checks of reference files	22
	Appendix II – Typical scores expected from ITU-T P.863 for a given codec.....	24
	Appendix III – Applications for further investigation	29
	Bibliography.....	30

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 [ITU-T P. 863].

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), <i>Vocabulary for performance and quality of service</i> ; plus its amendments. |
| [ITU-T P.863] | Recommendation ITU-T P.863 (2014), <i>Perceptual Objective Listening Quality Assessment</i> . |

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	Narrowband
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
SPL	Sound Pressure Level
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 (NB), wideband and super-wideband (SWB) 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 super-wideband 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:

1. the ability to compare results between different organizations;
2. 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] be run in parallel with [ITU-T P.863] until an appropriate feeling is developed for [ITU-T P.863] scores. Clause 12 provides additional information on the differences between [b-ITU-T P.862] and [ITU-T P.863].

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.

This 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:

1. narrowband: predicts quality as it would be perceived through an intermediate reference system (IRS) type receive filtered monaural handset;
2. super-wideband: 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 super-wideband 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 super-wideband 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 super-wideband mode be used?

The super-wideband mode should be used in all other cases (not covered by clause 7.2). Super-wideband mode is also appropriate for assessing narrowband connections. When assessing narrowband connections in super-wideband mode, a super-wideband 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 super-wideband 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 super-wideband scale. It is intended that [ITU-T P.863] super-wideband 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 super-wideband modes?

No. Although it is possible to apply an external IRS filter to the signals before passing to [ITU-T P.863] in super-wideband mode, the same score is not expected from the super-wideband 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 super-wideband mode.

7.6 Can I map a narrowband score to a super-wideband score?

No. The super-wideband 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 super-wideband mode will produce relatively lower scores for narrowband and wideband codecs compared to the scores observed in narrowband and wideband subjective tests.

The maximum super-wideband scale score of 4.75 was chosen to indicate that super-wideband 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] super-wideband scale has been engineered to enable easy comparison of signals with different bandwidths which may be difficult to achieve in subjective tests.