

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.862.3

(11/2007)

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

Methods for objective and subjective assessment of
quality

**Application guide for objective quality
measurement based on Recommendations
P.862, P.862.1 and P.862.2**

ITU-T Recommendation P.862.3

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Subscribers' lines and sets	Series	P.30
		P.300
Transmission standards	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 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

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

ITU-T Recommendation P.862.3

Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2

Summary

ITU-T Recommendation P.862.3 provides some important remarks that should be taken into account in the objective quality evaluation of speech conforming to ITU-T Recommendations P.862, P.862.1 and P.862.2. Users of ITU-T Recommendation P.862 should understand and follow the guidance given in this Recommendation.

This Recommendation forms a supplementary guide for users of ITU-T Recommendation P.862, which recommends a means of estimating listening speech quality by using reference and degraded speech samples. The scope of ITU-T Recommendation P.862 is clearly defined in itself. This Recommendation does not extend or narrow the scope, but provides necessary and important information for obtaining stable, reliable and meaningful objective measurement results in practice.

Source

ITU-T Recommendation P.862.3 was approved on 13 November 2007 by ITU-T Study Group 12 (2005-2008) under the ITU-T Recommendation A.8 procedure.

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 2008

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

CONTENTS

Page

1	Scope	1
2	References.....	1
3	Definitions	2
4	Abbreviations and acronyms	2
5	Conventions	3
6	General remarks.....	3
6.1	Testing factors	3
6.2	Applications.....	4
7	Characteristics of the reference signals	4
7.1	Length of signal.....	4
7.2	Active speech.....	5
7.3	Temporal structure.....	5
7.4	Active speech level.....	5
7.5	Application of artificial voice.....	5
7.6	Requirements for speech recordings	5
7.7	Variation in talker and speech content	6
7.8	Leading and trailing silences	6
7.9	Pre-filtering.....	6
7.10	Noise floor.....	6
7.11	Implementations issues.....	7
8	Characteristics of the degraded signal to be assessed.....	7
8.1	Difference in active speech duration between reference and degraded speech signal.....	8
8.2	Active speech level.....	8
8.3	Difference in duration of leading and trailing silence between reference and degraded speech.....	8
9	Characteristics of signal insertion and capturing paths	8
9.1	Influence of measurement circuits and test configuration in the insertion path	9
9.2	Influence of measurement circuits and test configuration in the capture path	10
10	Analysis of the results.....	10
10.1	Averaging the measurement results.....	10
10.2	Reliability of the PESQ measurements' results	10
10.3	Accuracy values of the PESQ measurements.....	11
10.4	Interpretation of the accuracy's results	12
11	Report of results.....	12
12	Guidance for using P.862.2 wideband extension to P.862	13



Documents

Document Sharing Hub

ITU-T Rec. P.862.3 (11/2007)

iii

	Page
13 Use of P.862.1 and P.862.2 use for EVRC type of codecs and evaluation of CDMA networks.....	13
14 Comparing objective with subjective score.....	14
Appendix I – Reference values for objective quality derived by ITU-T Recommendation P.862 for ITU-T/GSM standard codecs.....	15
I.1 Reference value sources	15
I.2 Pre-processing of source speech.....	17
I.3 Processing of G.711.....	18
I.4 Processing of G.726.....	18
I.5 Processing of G.728, G.729, Annex A/G.729 and G.723.1	18
I.6 Processing of MNRU	18
Appendix II – Test databases for P.862/P.862.1	22
Appendix III – Report of P.862/P.862.1 measurements	23
III.1 Report and interpretation of the average PESQ results	23
III.2 Report and interpretation of individual PESQ measurement results.....	23
Appendix IV – Calibration method for proprietary interfaces.....	25
IV.1 Calibration of the transmit level (near end) of the test equipment	25
IV.2 Calibration of the receive level (far end) of the test equipment	25
Bibliography.....	26

ITU-T Recommendation P.862.3

Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2

1 Scope

This Recommendation provides some important remarks that should be taken into account in the objective quality evaluation of speech conforming to [ITU-T P.862], [ITU-T P.862.1] and [ITU-T P.862.2]. Users of [ITU-T P.862] should understand and follow the guidance given in this Recommendation.

This Recommendation forms a supplementary guide for users of [ITU-T P.862], which recommends a means of estimating listening speech quality by using reference and degraded speech samples. It cannot be used for the assessment of talking quality or interaction quality. It assumes that an objective quality estimation algorithm strictly conforms to [ITU-T P.862]. This can be confirmed by the conformance test provided as an annex to [ITU-T P.862].

The scope of [ITU-T P.862] is clearly defined in itself. This Recommendation does not extend or narrow the scope, but provides necessary and important information for obtaining stable, reliable and meaningful objective measurement results in practice.

Applications and limitations associated with the wideband extension to P.862 defined in [ITU-T P.862.2] are discussed in clause 12.

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.50] | ITU-T Recommendation P.50 (1999), <i>Artificial voices</i> . |
| [ITU-T P.56] | ITU-T Recommendation P.56 (1993), <i>Objective measurement of active speech level</i> . |
| [ITU-T P.501] | ITU-T Recommendation P.501 (2007), <i>Test signals for use in telephony</i> . |
| [ITU-T P.800] | ITU-T Recommendation P.800 (1996), <i>Methods for subjective determination of transmission quality</i> . |
| [ITU-T P.830] | ITU-T Recommendation P.830 (1996), <i>Subjective performance assessment of telephone-band and wideband digital codecs</i> . |
| [ITU-T P.862] | ITU-T Recommendation P.862 (2001), <i>Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs</i> . |
| [ITU-T P.862.1] | ITU-T Recommendation P.862.1 (2003), <i>Mapping function for transforming P.862 raw result scores to MOS-LQO</i> . |
| [ITU-T P.862.2] | ITU-T Recommendation P.862.2 (2007), <i>Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs</i> . |

3 Definitions

This Recommendation defines the following terms:

3.1 source speech/signal: The original speech signal without any degradation. This should be recorded and stored conforming to [ITU-T P.830]. It may or may not be the same as reference speech defined below.

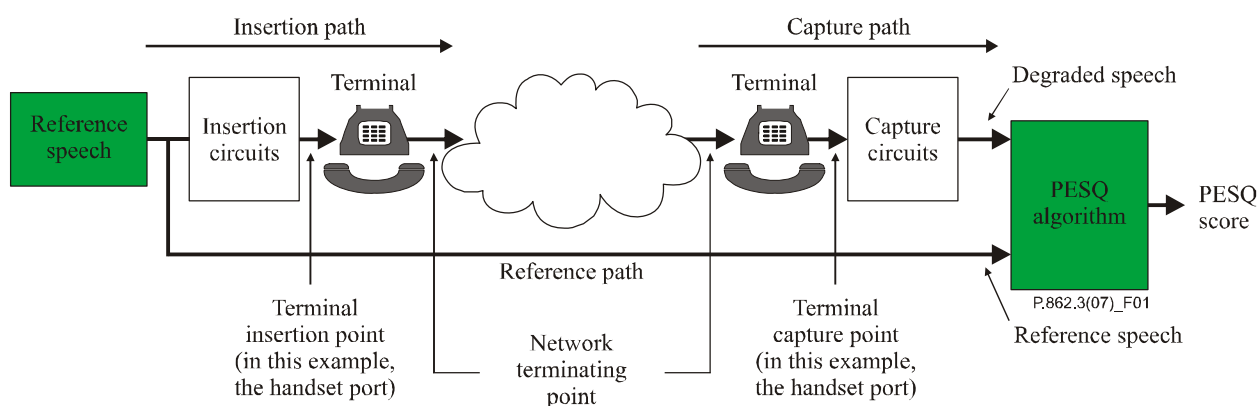
3.2 reference speech/signal: The speech signal to be used by the [ITU-T P.862] algorithm as a reference against which the effects of the system under test are revealed.

3.3 input speech/signal: The signal fed into the system under test at the signal insertion point. It is derived from the reference speech signal. It may be identical to the reference signal or it may be processed by e.g., overlaying it with noise. Further information is provided in clause 7.10.

3.4 degraded speech/signal: The reference speech that has passed through the system under test.

3.5 signal insertion path: Consists of the connection path (wiring, electronics, etc.) between the reference signal to the [ITU-T P.862] algorithm and the input interface, called the insertion point.

Figure 1 divides an example test circuit into insertion path, system under test, and capture path, and shows possible insertion and capture points *in the case of hardware measurements*. The particular insertion and capture points will depend on the specific system under test and the configuration of the test set-up.



NOTE – Depending on how the signals are inserted, captured and stored, and whether the ITU-T Rec. P.862 measurement computations are done in real-time or deferred, these paths may be continuous physical electrical pathways, or they may be logical paths, such as when an output sample is stored for later analysis.

Figure 1 – Example of measurement set-up and terminology

3.6 signal capture path: Consists of the connection path between the capture point (output interface with the network under test) and the [ITU-T P.862] algorithm (see Figure 1).

3.7 dBov: The value in dB relative to the overload point of a digital system. According to ITU-T Recommendation G.711, 0 dBm0 in analogue representation corresponds to –6.15 dBov and –6.18 dBov for A-law and μ -law codecs, respectively.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

IRS	Intermediate Reference System
MOS	Mean Opinion Score

MOS-LQO	Mean Opinion Score – Listening Quality Objective (an estimation of subjective listening quality using an objective measurement technique)
MOS-LQS	Mean Opinion Score – Listening Quality Subjective (a direct measurement of listening quality using subjective ratings of samples)
PESQ	Perceptual Evaluation of Speech Quality
RMS	Root Mean Square

5 Conventions

It is recommended that raw results from [ITU-T P.862] measurements be converted to MOS-LQO (as defined in ITU-T Recommendation P.800.1) using the relation defined in [ITU-T P.862.1.]¹. This will prevent potential confusion in comparison and interpretation of results due to the superficial similarity of the two scales.

6 General remarks

This clause gives supplementary remarks about the scope of [ITU-T P.862]. The scope itself is summarized in [ITU-T P.862] quite clearly.

The reliability and consistency of the results are dependent on several factors, for example:

- Number of calls.
- Number of measurements.
- Length of speech samples.
- Type of speech used, e.g., natural or artificial.

These factors, and the following considerations, affect the structure and complexity of the tests:

- Purpose of the measurements (e.g., for benchmarking connections, routine monitoring or fault diagnosis).
- Transmission channel characteristics (e.g., do the channel characteristics vary over time such as with mobile or some types of VoIP connection?).
- Time available to make the measurements (which may impact on the scope of the testing).

Where it is suspected that certain connection types may be affected by 'busy-hour' conditions, it may also be important to carry out a number of sequences of measurements at different times throughout the day.

The test structure used should always be quoted in conjunction with processed values of MOS-LQO.

6.1 Testing factors

[ITU-T P.862] is validated for the evaluation of test factors, coding technologies and applications, which are listed in Table 1 of [ITU-T P.862]. In particular, care should be taken when one carries out live network testing since there might be some equipment that causes degradation that [ITU-T P.862] cannot handle, e.g., artefacts caused by noise reduction systems, in between the signal insertion point and signal capture point. It is also known that PESQ underestimates severe linear frequency response distortions. This applies especially to, e.g., bandwidth limitations narrower than 300 Hz ... 3.4 kHz.

¹ The detailed procedure for obtaining MOS-LQO can be found in clause 10.

Use of [ITU-T P.862] with systems that include noise suppression algorithms between the signal insertion point and the signal capture point is not recommended.

6.2 Applications

[ITU-T P.862] can be used as a means for live network testing, in which one evaluates the system under live conditions rather than computer-simulated conditions or fixed test set-up in a laboratory environment.

Live field testing will not produce repeatable results due to uncontrolled time-varying transmission channels. The alternatives are controlled network simulations with exactly repeatable results. For the latter condition, averaging should be used.

Live field network testing, such as mobile drive testing, will affect the structure and content of the reference speech signals. In drive testing this is due to the necessity to assess a very highly time-varying quality in order to get accurate geographical quality information.

Live field network testing also presents the need to assess quality on a per-sample basis since per-condition averaging is not possible with live, time-varying network conditions.

Both of the above will have an effect on the stability and possibly on the accuracy of [ITU-T P.862] results. For this reason, users of [ITU-T P.862] in live network testing should be careful to check results and repeat measurements with a view to checking result stability. Performance results are shown in clause 10.2.

If the system under test involves a broadband terminal (such as certain hands-free headsets, or broadband IP phones) then PESQ will predict the quality as it would have been perceived for IRS-type receive filtering.

7 Characteristics of the reference signals

Reference signals are defined and used as input signals to the system under test and as the reference input for [ITU-T P.862]. The characteristics of the signal insertion path are discussed separately in clause 9. If the language under evaluation is included in the speech database provided as Annex B of [ITU-T P.501], we recommend using it as test signals to improve the compatibility among different measurements by avoiding the use of different reference signals.

7.1 Length of signal

[ITU-T P.862] has been validated in ITU-T for use with signals that are mostly 8-12 s long. However, it is known that [ITU-T P.862] can be applied to speech up to 30 s long [b-ITU-T COM12-D008]. Therefore, it is recommended that each speech sample should be 8-30 s long. This includes any silence before, after and between utterances².

For live field test scenarios, shorter reference signals may be used, however this may not exercise the system as fully as possible. These shorter sentences should use at least the 3.2 s of speech as defined in clause 7.2.

² The reference software provided as Annex A to [ITU-T P.862] has the following limitation with respect to the length of signals, although this limitation is already out of the range determined in this Recommendation: Due to the precision available to the floating point arithmetic in [ITU-T P.862], once the signals being processed reach a certain length, errors will start to be introduced in the signal energy calculation. Analysis suggests that signals with more than about one million samples will start to cause problems. Sixty seconds of a 16 kHz monoaural signal contains 960'000 samples and this would be a sensible threshold at which to apply a warning.

It should be noted that, because of the non-linearity of the [ITU-T P.862] algorithm, the result obtained using concatenated signals will not correspond to the simple arithmetic mean of the results for individual samples.

7.2 Active speech

The speech activity in the reference speech, which can be measured based on [ITU-T P.56]³, should be between 40% and 80%. There should be a minimum of 3.2 s active speech in the reference. In combination with the recommended signal file length, this should ensure that [ITU-T P.862] has enough speech to make an accurate prediction and the speech should contain some silence to exercise important elements in the network.

7.3 Temporal structure

Reference speech should comprise utterances separated by silent periods representative of natural pauses in speech. Most of the experiments used in calibrating and validating [ITU-T P.862] contained pairs of sentences separated by silence. Good examples are speech materials included in [ITU-T P-series Supp.23]⁴, which last 8 s and include two short sentences separated by a silent period of at least 1 s, and in Annex B of [ITU-T P.501] as mentioned above. It is recommended that the reference speech includes a few continuous utterances rather than many short utterances of speech such as rapid counting⁵.

7.4 Active speech level

The active speech level referred to in this Recommendation is the equivalent level of the digitally stored reference signal, as measured according to [ITU-T P.56]. The active speech level applied to the signal insertion path of the measurement system is separately described in clause 9. It is recommended that all the reference speech files be stored at a level of –30 dBov to avoid peak clipping. Note that this is the level of source speech stored in the digital format and that the input level to the system under test should be determined separately according to the purpose of objective measurement (see [ITU-T P.830]).⁶

7.5 Application of artificial voice

The application of artificial voice signals needs more investigation, from the viewpoints of language and temporal structure of signal power, and possibly other factors [b-ITU-T COM12-D145].

7.6 Requirements for speech recordings

[ITU-T P.800] and [ITU-T P.830] give guidance for recording speech materials. This Recommendation assumes that source speech is recorded in conformance with this guidance. Note

³ ITU-T Recommendation G.191 provides a software called sv56demo.c, which measures the active speech ratio and active speech level conforming to [ITU-T P.56].

⁴ Please note that the copyright on [ITU-T P-series Supp.23] does not allow the use of the signals in commercial applications.

⁵ The reference software provided as Annex A to [ITU-T P.862] has a limit of 50 as the maximum number of utterances. If reference signals with many utterances are used, it must be verified that the implementation of [ITU-T P.862] used for the test can handle that large a number.

⁶ A typical nominal value for active speech level is –20 dBm0, corresponding to approximately –26 dBov. In any specific system to be tested, the mean active speech level in the system under test may be significantly different from the nominal value of –20 dBm0. In such cases, the measured mean value may be used as the input active speech level. When the system response to input level is being assessed, it is appropriate to use a range of active speech values, for example, –14, –26 and –38 dBov (approximately equivalent to –8, –20 and –32 dBm0) as recommended by [ITU-T P.830].