International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.862.3

(11/2005)

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

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

This Recommendation forms a supplementary guide for users of ITU-T Rec. P.862, which recommends a means of estimating listening speech quality by using reference and degraded speech samples. The scope of ITU-T Rec. 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 29 November 2005 by ITU-T Study Group 12 (2005-2008) under the ITU-T Recommendation A.8 procedure. This text includes the clarifications agreed on 13 June 2006 by ITU-T Study group 12.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. 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, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

Page

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 Recs P.862, P.862.1 and P.862.2. Users of ITU-T Rec. P.862 should understand and follow the guidance given in this Recommendation.

This Recommendation forms a supplementary guide for users of ITU-T Rec. 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 Rec. P.862. This can be confirmed by the conformance test provided as an annex to ITU-T Rec. P.862.

The scope of ITU-T Rec. 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 Rec. 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.

[1] ITU-T Recommendation P.501 (2000), *Test signals for use in telephonometry, Annex B – Speech files and noise sequences*.

[2] ITU-T Recommendation P.56 (1993), *Objective measurement of active speech level*.

[3] ITU-T P-series Recommendations – Supplement 23 (1998), *ITU-T coded-speech database*.

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

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

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

[7] ITU-T Recommendation 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*.

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

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 Rec. P.830. It may or may not be the same as reference speech defined below.

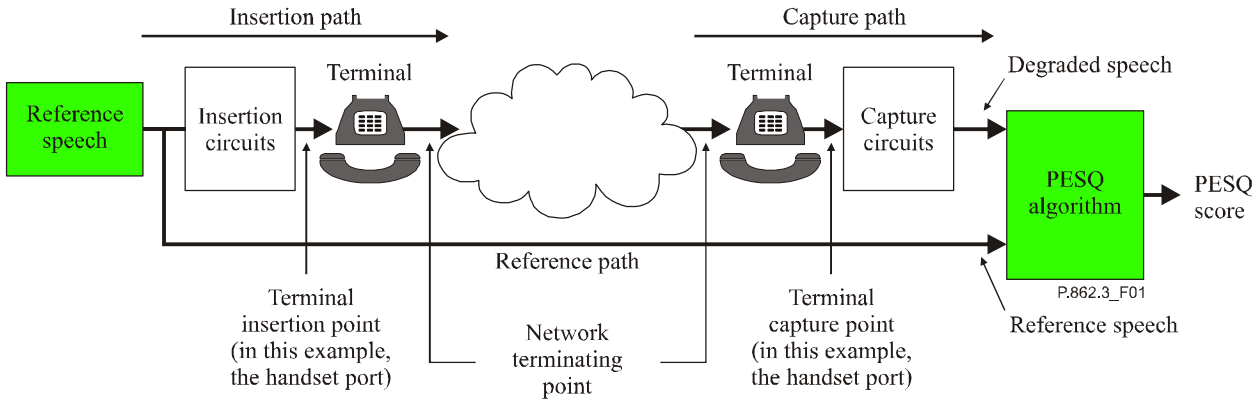
3.2 reference speech/signal: The speech signal to be used by the ITU-T Rec. P.862 algorithm as a reference against which the effects of the system under test are revealed.

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 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 Rec. 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/P.862.3 – 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 Rec. 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 Rec. 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

This Recommendation uses the following abbreviations:

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 Rec. P.862 measurements be converted to MOS-LQO (as defined in ITU-T Rec. P.800.1) using the relation defined in ITU-T Rec. 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 Rec. P.862. The scope itself is summarized in ITU-T Rec. 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 Rec. P.862 is validated for the evaluation of test factors, coding technologies and applications, which are listed in Table 1/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 Rec. P.862 cannot handle, e.g., artefacts caused by noise reduction systems, inbetween 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 Rec. 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 Rec. 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 Rec. P.862 results. For this reason, users of ITU-T Rec. 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 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 Rec. 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/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 Rec. 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 Rec. P.862 can be applied to speech up to 30 s long [B.1]. Therefore, it is recommended that each speech sample should be 8-30 s long. This includes any silence before, after, and between utterances[2].

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 7.2.

---

[2] The reference software provided as Annex A/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 Rec. 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 Rec. 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 Rec. 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 Rec. 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 Rec. P.862 contained pairs of sentences separated by silence. Good examples are speech materials included in Supplement 23 to P-series Recommendations[4], which last 8 s and include two short sentences separated by a silent period of at least 1 s, and in Annex B/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 counting[5].

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 Rec. 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 Rec. 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.2].

7.6 Requirements for speech recordings

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

---

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

[4] Please note that the copyright on Supplement 23 does not allow the use of the signals in commercial applications.

[5] The reference software provided as Annex A/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 Rec. 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 Rec. P.830.

speech may be the same as this source speech or it may have added low-level noise floor and/or frequency shaping (see 7.9 and 7.10).

7.7 Variation in talker and speech content

Variation due to the talker and speech content can be controlled by using a fixed set of samples for all test cases to be compared. Therefore, it is helpful to use the speech database provided as Annex B/P.501 to facilitate *post-hoc* comparisons and interpretation of results from different laboratories.

For network simulation scenarios it is recommended that the reference speech should include a minimum of two female and two male talkers, each speaking different sentences. The P.862.1 scores obtained with these different samples should be averaged for a per-condition evaluation afterwards

For live field test scenarios, less speaker variation may be used, however this may not exercise the system as fully as possible. If this format is necessary, multiple speakers might be included in these short reference signals. In case of an intended per-sample evaluation, samples containing more than one speaker's voice will decrease the sample dependency of the derived results.

During the validation of ITU-T Rec. P.862, very little data was available for children's voices and certain speech characteristics (e.g., voice/speech disorders, etc.). With the limited data available, no problems were observed with children's voices. Music must not be used with ITU-T Rec. P.862.

It is also recommended to use several different speech samples (4-10 sentences) per talker to reflect phonetic variations.

7.8 Leading and trailing silences

ITU-T Rec. P.862 uses the RMS level of the reference and degraded signals for level alignment. If long silences are included at the beginning and end of the reference signal, then the level alignment result may be compromised.

A minimum leading and trailing silence of 0.5 s is recommended, as long as the measurement equipment can synchronize the degraded speech with the reference one within that time.

A maximum leading and trailing silence of about 2 s is recommended and may be useful if there is a high level of delay in the system.

7.9 Pre-filtering

The reference speech prepared according to 7.1 to 7.8 should be filtered so that the sending frequency characteristics of a handset are taken into account. It should be noted that ITU-T Rec. P.862 assumes that reference speech reflects such electro-acoustic characteristics appropriately. When one assumes that the reference speech is fed into networks as the output of a handset terminal, ITU-T recommends the use of the modified IRS sending characteristics defined in Annex D/P.830. Such filtering should be done after 7.1 to 7.8 have been taken into account appropriately.

Care should be taken to coordinate the filtering used with the nominal frequency response of the system under test, because such filtering is dependent on where one feeds the reference speech to equipment and/or networks under test (see clause 9).

7.10 Noise floor

The noise floor in reference speech should be adequately low as expected in recordings conforming to ITU-T Recs P.800 and P.830. It is also possible to add complete silence (e.g., a signal having a digital amplitude of zero) so that the reference speech signals have the proper characteristics

defined in 7.1, 7.2, 7.3 and 7.8[7]. This is the case where the reference speech corresponds to the source speech, as described in 7.6.

If one anticipates unwanted noise in the measurement paths described in clause 9 or the noise floor in the device under test itself, however, a low noise floor of about –75 dBov, white spectrum, should be intentionally added to the reference signal as mentioned above and stored in the 16-bit linear PCM format. The level of the noise floor should be determined within 0-4000 Hz[8]. Noise at this level will not adversely affect the results based on ITU-T Rec. P.862, but will effectively remove the contribution of such measurement noise to the final score [B.3]. It is quite important to add such a noise floor after to the pre-filtering described in 7.9.

The active speech level at the signal insertion path of the measurement system described in clause 9 should be calibrated after such pre-filtering[9].

It should be noted that the proposed insertion of additional noise into the reference signal will lead to more accurate results if the unwanted noise at the receiving path is a continuous noise-floor and it does not solve problems coming up with comfort noise, which is only inserted in speech pauses.

7.11 Implementations Issues

Many signals included in the P.862 conformance test do not fulfil the requirements set forth above. For the conformance test this does not matter at all since the sole purpose is to prove the correctness of the implementation. Care must however be taken that the implemented algorithm also produces results in cases of violation of the requirements defined in this Recommendation since otherwise the conformance test cannot be applied.

8 Characteristics of the degraded signal to be assessed

Degraded signals are the output of the system under test that correspond to the test input, including any effects due to the measurement interface. This clause describes the characteristics of signals

---

[7] Sending a digital silence into a digital insertion circuit, e.g., ISDN phone, and then to a wireless phone link may cause undesirable side effects if GSM or 3GPP codecs are used in the wireless network. Namely, sending a constant pattern of the lowest positive G.711 A-law value of +8 linear (D5 PCM hex) will cause the speech codec to reset at a period of 20 ms, i.e., 50 times/s. This is due to the built-in codec homing procedure for codec testing purposes. Although the resetting during speech pauses is not harmful for the codec itself, measured speech quality may not be the same compared with the normal situation in which codecs are not reset. A side effect would be that the codec would not use the discontinuous transmission (DTX) during speech pauses although this is enabled by the network. Therefore, the comfort noise effect or possible speech clipping by the voice activity detector (VAD) would not be verified at all. To overcome this specific problem, a low-level noise floor of approximately –65 dBm0 should be added to the test sample before it is sent to the network. This will break the constant pattern of digital silence.

[8] When one employs 16 kHz as the sampling rate of reference speech, care should be taken in determining the level of the noise floor.

[9] The noise floor may not be fed into the system under test because the level of the noise floor becomes less than the minimum possible level of the system. For example, the lowest values for A-law coding are ± 8 in linear 16-bit values. Thus, the lowest level is −72 dBov. This means that, if the input level to the system is calibrated to −30 dBov for instance, the noise floor at –75 dBov cannot go through the system and does not solve the unwanted noise problem at all. If significantly higher input levels than a nominal active speech level of –26 dBov is to be tested, and if a stored noise floor of –75 dBov is applied to the reference sample, possible degradations (e.g., clicks, noise bursts, etc.) on speech pauses may not be estimated by ITU-T Rec. P.862, because a higher noise floor level in the degraded sample may mask low level degradations. By using a nominal active speech level of approximately –26 dBov, however, this problem does not exist.

that are digitally stored as output of the system under test for use in the calculation based on ITU-T Rec. P.862. The characteristics of the signal capture path are discussed separately in clause 9.

8.1 Difference in active speech duration between reference and degraded speech signal

The active speech duration is defined in ITU-T Rec. P.56.

ITU-T Rec. P.862 uses the RMS levels of the reference and degraded signals for level alignment. This means that the algorithm may give erroneous results if speech is missing or if silence is added to, or taken away, from the degraded signal.

When an utterance has been deleted from the degraded signal, or if one or more large sections of the degraded signal have been muted, the signal will be level-shifted to a value above the actual value.

When silence has been taken out of the degraded signal, the signal will be level-shifted to a value below the actual value.

These concerns will affect the amount of disturbance present in the degraded signal and will therefore affect the objective quality measurement result. If the durations of speech in the reference and degraded signals differ by more than 25%, the effect may be large enough to significantly bias the result. This is especially true if long continuous sections of speech have been replaced by silence.

8.2 Active speech level

The active speech level is defined in ITU-T Rec. P.56.

Although the active speech level is normalized in calculating PESQ values, it is recommended that the digital speech level stored as degraded signals to the PESQ algorithm should be around –30 dBov to avoid clipping and quantization distortion. It should be noted that ITU-T Rec. P.862 cannot be used to evaluate the effects of the receiving/listening level[10].

8.3 Difference in duration of leading and trailing silence between reference and degraded speech

ITU-T Rec. P.862 uses the RMS level of the reference and degraded signals for level alignment. If long pauses are included at the beginning and end of the degraded signal, then the level alignment process may be sub-optimal. This issue may become a problem if the reference and degraded signal durations differ by more than 20%[11].

Additionally, ITU-T Rec. P.862 does not take into account any distortion in the degraded signal occurring before the start or after the end of the active speech signal. This active speech signal is determined from the reference signal as the first and last points where the signal level goes above approximately 50 dB SPL.

9 Characteristics of signal insertion and capturing paths

This clause describes the desired characteristics of the signal insertion and capturing paths in hardware measurements. The measurement circuitry and environment can affect ITU-T Rec. P.862

---

[10] If an MNRU conforming to ITU-T Rec. P.810 is being tested, care should be taken in the level equalization process to preserve the actual speech level excluding the noise added by the MNRU.

[11] Empirical observations suggest that ITU-T Rec. P.862 results for EVRC [B.4] depend on the particular alignment of the coding frame boundaries with the input PCM data. The result may vary by up to 0.25 depending on where the frame boundaries fall. In the case of EVRC, the method of obtaining a stable result would be to measure each of the 80 possible alignments and average the results. Similar situations may be uncovered for other DSP processes.

results unless precautions are taken to control the factors involved. Noise and interference must be excluded from the insertion and capture paths as much as possible to ensure that they do not affect the results.

9.1 Influence of measurement circuits and test configuration in the insertion path

If the possibility exists to use a well-defined interface like POTS or ISDN, than this is the preferred method and the test equipment should be calibrated to serve such interfaces with the recommended nominal signal levels.

If no such well-defined interface can be used, then the insertion point is often the handset port of an end device, which is a proprietary interface and the required input level is initially unknown. Although standards such as North America's TIA-810-A specify terminal characteristics between acoustic and network interfaces, they do not specify intermediate points such as the handset interface. Gain distribution and filtering will be specific to the vendor, or even the individual terminal. In some cases, these characteristics are configurable in the end device. When the handset port is used as an ITU-T Rec. P.862 test port, the test engineer may intend to measure:

1) the performance of the end device and network together;

2) the performance of the end device by itself (connected to a reference network); or

3) the performance of the network itself, with minimal contribution from the end device.

In all cases, however, we want to eliminate contributions from the measurement set-up.

The test engineer should ensure that the active speech level applied to the proprietary interface is consistent with the desired network level and the dynamic range of the codec. This requires appropriate gain characterization between the insertion point and the interface point of the terminal and the network in both directions of transmission.

When applying speech signals to the proprietary interface, the test engineer should be aware of filtering (e.g., modified IRS and frequency equalization of the transducer) between the acoustic and network interfaces. Terminal vendors are free to implement any combination of acoustic, electronic, or digital filtering on either side of the proprietary handset interface. P.862 test equipment may therefore see either complete, partial, or no filtering after the insertion point. To obtain a precise result, the configuration measured must include the filtering appropriate for the test case being observed. Likewise, the filtering applied to the reference signal must match the filtering applied in the end-to-end test circuit.

This paragraph explains an ideal technique that can be used in order to determine the input characteristic for the reference signal when there is the possibility that an input filter used in normal operation of the communication terminal device is not in the measurement path. An artificial mouth (for example that found on a Head and Torso Simulator) should be used to inject the test signal acoustically into the terminal which should be connected to a far-end reference point (e.g., ISDN point). The acoustic level used should represent normal usage for the terminal device, and the background noise level should be below 35 dBA. This normal use may either reflect the usage of the internal microphone or a personal hands-free kit, and depends on the purpose of the scenario to be assessed. The artificial mouth should be calibrated, and the positioning of the terminal should be representative of normal usage. The electrical level and frequency value should be measured at a reference point in the network connection (e.g., the ISDN end point). The process should then be repeated (with the same test signal) using electrical input at the P.862 test injection point, using the equipment used during P.862 testing. The input signal should now be adjusted so that the electrical level and frequency value match those captured during acoustic injection. The technique described

here is the ideal method, and may be approximated for many situations[12]. If this technique is not used, it is recommended that the tester takes special notice of manufacturers' specifications for acoustic and electrical interfaces to the communication terminals.

9.2 Influence of measurement circuits and test configuration in the capture path

Once the reference speech has passed through the system under test, it must be transferred from the capture point to P.862. This capture path can contribute noise and distortion which may affect the result. The capture path may be subject to difficulties such as ground loops, pickup from a.c. power conductors, or other common-mode signals that may be present. In-band pickup may bias the result. In addition, out-of-band noise at sufficiently high levels may alias into the measurement band where insufficient anti-aliasing filtering is used.

To minimize noise contributed by the insertion and capture paths, it is recommended that insertion and capture paths together should contribute less than –70 dBov, so that the resultant SNR becomes 40 dB and the objective quality measurement result is determined exclusively by the influence of the system under test.

In general slowly varying sampling rates, time stretching or time compression of the transmitted signal may lead to too pessimistic scores due to improper time alignment.

If analogue transmission is involved care must be taken that no excessive clock drifts between the AD and DA converters occur. This may be the case with consumer equipment, especially if the hardware does not support the required sample rate and a software sample rate conversion by the driver of the sound card is involved.

10 Analysis of the results

10.1 Averaging the measurement results

As highlighted in 7.7, one should use at least two female and two male talkers in an objective measurement. Before computing the mean or other statistics, individual measurement results should first be transformed to the MOS-LQO domain (based on ITU-T Rec. P.862.1) and then averaged over talkers and speech samples. Since the algorithm defined in ITU-T Rec. P.862 is non-linear, results from concatenated samples will not match the mean results from those samples tested individually.

As mentioned in 6.2, there are two types of P.862 application, which require different analysis approaches. In the first case, averaging over talkers and speech samples should be performed before continuing with the analysis. This analysis is suitable for controlled network simulations with exactly repeatable results. The case of live field network testing needs a per-sample quality evaluation, due to uncontrolled time-varying transmission channels.

10.2 Reliability of the PESQ measurements' results

A large number of databases have been used for P.862 testing, validation and calibration (P.862.1). As described in ITU-T Recs P.862 and P.862.1, the databases contained speech samples spoken by different talkers and genders, in different languages and representing speech degradations generated by simulated and live network conditions. In addition, the network conditions corresponded to fixed, wireless and VoIP applications. Details regarding the content of the test databases are presented in Appendix II.

---

[12] This approximation is measurement-equipment specific. One possible method is described in Appendix IV.

It should be noted that the P.862/P.862.1 measurement results are 95% reliable, and exhibit a known and controlled accuracy, when the algorithm is used on the same type of applications as the ones on which the algorithm has been trained, tested and validated. In other words, the measurement scenarios need to represent statistically the same type of sample population as the ones on which P.862/P.862.1 has been trained, tested, validated and calibrated in order for the determined accuracy values to remain valid. The results' reliability and accuracy become unknown and uncontrolled once the algorithm is used to evaluate speech quality on new types of technologies and/or using other types of codecs and/or new live networks.

10.3 Accuracy values of the PESQ measurements

Three statistical metrics, the correlation coefficient, the prediction error and the residual error distribution, have been used to evaluate P.862/P.862.1 performance on the databases described in 10.2. As mentioned in 10.1, the analysis approaches differ depending on the application type, controlled network simulations and live/field network testing conditions.

For all simulated network conditions, averages per-condition over at least four talkers, 2 male and 2 female, have been used to calculate the statistical metrics. For live network databases, the statistical metrics have been calculated using per-samples objective and subjective scores.

The performance results are presented in Tables 1 and 2. The 95% confidence critical limits for the correlation coefficient and the prediction error are also calculated in order to provide the 95% lower correlation bound and 95% upper prediction error bound.

The results are presented per application type (e.g., simulated wireless and VoIP network conditions and real-life wireless and VoIP network conditions). These accuracy values express therefore PESQ algorithm's performance if used in any of the applications mentioned in 10.2.

Table 1/P.862.3 – Confidence intervals for correlations coefficient and prediction error

Application	N	Metric	P.862 (raw PESQ)	P.862.1 (calibrated PESQ)
Simulation data (wireless, VoIP and fixed applications)	1357	R	0.956	0.956
		CI95%-lower limit	0.940	0.940
		PE	N.A.	N.A.
		CI95%-upper limit	N.A.	N.A.
Field collected data (Wireless applications: GSM US and EU, CDMA-US, TDMA-US, iDEN-US, AMPS-US; and VoIP application)	1135	R	0.925	0.926
		CI95%-lower limit	0.916	0.917
		PE	0.479	0.462
		CI95%-upper limit	0.492	0.475

Table 2/P.862.3 – Residual error distribution

Application	MOS bins	<0.25	<0.5	<0.75	<1	<1.25	<1.5	<1.75	<2
Field collected data (Wireless applications: GSM US and EU, CDMA-US, TDMA-US, iDEN-US, AMPS-US; and VoIP application)	P.862 CDF (%)	32.51	66.52	90.84	97.97	99.38	99.91	99.91	100
	P.862 prob (%)	32.51	34.09	24.32	7.14	1.41	0.53	0	0.09
	P.862.1 CDF (%)	40.44	70.48	90.33	97.71	99.3	99.7	99.91	100
	P.862.1 Prob (%)	40.44	30.04	19.82	7.4	1.59	0.44	0.18	0.09

10.4 Interpretation of the accuracy's results

By definition, the P.862/P.862.1 algorithm is an estimator of the subjective opinion on the speech quality provided by the network under test. It should be noted therefore that any speech quality measurement performed by the PESQ algorithm is affected by the accuracy values presented in Tables 1 and 2.

It should be noted that, as mentioned in 10.2, the accuracy values remain valid as long as the measurement scenarios represent statistically the same sample population as the ones presented in 10.2.

The lower bound of the 95% confidence interval of the correlation coefficient shows that P.862/P.862.1 measurements are expected to exhibit a correlation with the subjective opinion, which is higher or at least equal to the lower limit of the correlation coefficient 95% confidence interval, regardless of whether simulated or live field network conditions are used, and regardless of the tested network type (such as wireless, VoIP and fixed) (Table 1).

The residual error distribution (Table 2) represents the cumulative density function (CDF) of the absolute errors between MOS and P.862/P.862.1 scores and it shows the probability that the absolute error is lower than a value. For example, the probability that the absolute error is lower than 0.5 MOS is higher than 70%, while the probability that the error is lower than 0.75 MOS is higher than 90%. Table 2 also provides the probability density function (PDF) of the absolute error. As expected, in agreement with the CDF, the PDF shows that lower absolute errors have a higher likelihood of occurrence than higher values.

11 Report of results

As mentioned in 10.4, depending on the application type, i.e., simulated or live network conditions, the P.862/P.862.1 measurements should be reported based on the algorithm's accuracy presented in 10.3 (Tables 1 and 2).

The correlation coefficient is recommended to be used as an informative statistical metric on the P.862/P.862.1 performance for a specified application. The prediction error along with the residual error distribution is recommended to be used to report P.862.1 measurement results for a specified application.

Generally, average, maximum, and minimum PESQ values should be reported, as well as the number of measurements used to calculate the average. Some detailed recommendations for reporting PESQ measurements are presented in Appendix III. In addition, the number of measurements achieving a given PESQ score or range of scores can be presented graphically as a frequency distribution. In cases where the system under test delivers relatively stable listening

quality, the standard deviation can be used to help decide whether further measurements are necessary to achieve a specified accuracy. This approach is not valid for highly time varying systems under test (e.g., VoIP or mobile networks).

12 Guidance for using P.862.2 wideband extension to P.862

In principle, the guidance provided in this Recommendation is applicable to both ITU-T Rec. P.862 and its wideband extension P.862.2. Nevertheless, some specific guidance is necessary for the wideband extension to ITU-T Rec. P.862.

The foregoing guidance is mainly referring to the usage of the IRS send characteristic to be applied on the input or reference signal. For the wideband extension, no filtering of either the speech signal or any environmental noise is recommended. This is referred to in 6.2, 7.9 and 7.10.

Regarding speech activity level calculation according to ITU-T Rec. P.56, it is recommended to use the P.56 wideband option. This is referred to in 7.2, 7.4, 8.1 and 8.2.

The proposed insertion of a low noise floor in 7.10 is not evaluated for the wideband extension and cannot be recommended.

Clauses 10 and 11 describe the accuracy of the P.862 method. The figures provided are only applicable to the narrow-band P.862.

Within 3.6 the 0 dBm0 is described according to ITU-T Rec. G.711. It has to be stated that this G.711 reference is only available for narrow-band applications.

Both methods, P.862.1 and P.862.2, refer to a MOS-LQO scale as defined in ITU-T Rec. P.800.1. It has to be taken into account that the term MOS-LQO might be extended by a qualifier relating to the narrow-band and wideband cases in the future. Please note that the results produced by ITU-T Rec. P.862.1 are related to a narrow-band-only context. ITU-T Rec. P.862.2 results apply to wideband or mixed wideband and narrow-band applications. As a result, direct comparisons of the P.862.1 and P.862.2 MOS-LQO results are not possible.

Appendix I

Reference values for objective quality derived by ITU-T Rec. P.862 for ITU-T/GSM standard codecs

I.1 ITU-T Rec. P.862.1 reference values were calculated for the following codec/MNRU conditions by using the speech database in Annex B/P.501:

- G.711 μ-law, A-law;
- G.726 at 16, 24, 32, and 40 kbit/s;
- G.728;
- G.729;
- Annex A/G.729 ;
- G.723.1 at 5.3 and 6.3 kbit/s;
- GSM-AMR at 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, and 12.2 kbit/s;
- GSM-EFR;
- GSM-FR;
- GSM-HR;
- MNRU (Q = 5, 10, 15, 20, 25, 30, 35, 40 and 45 dB).

Figure I.1 shows the preprocessing of the reference speech signals. Figures I.2, I.3, and I.4 show the coding procedures of G.711, G.726, and the other codecs, respectively. Figure I.5 shows the processing procedure of MNRU. Tables I.1, I.2, and I.3 show reference values, which were derived by transforming raw PESQ values[13] to MOS-LQO by using ITU-T Rec. P.862.1, for ITU-T standardized codecs and MNRU. Signal processing of G.711, G.726, and MNRU was done using the software tool provided in ITU-T Rec. G.191.

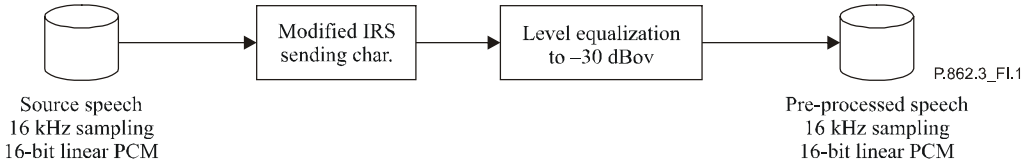


Figure I.1/P.862.3 – Pre-processing of source speech

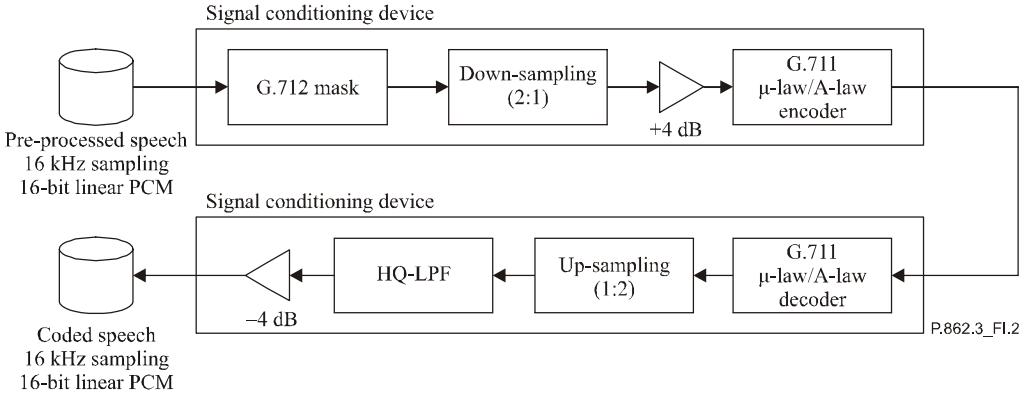


Figure I.2/P.862.3 – Processing of G.711

[13] In GSM-AMR, GSM-EFR, GSM-FR, GSM-HR, and G.711 A-law codecs, raw PESQ values were calculated at a sampling rate of 16 kHz.

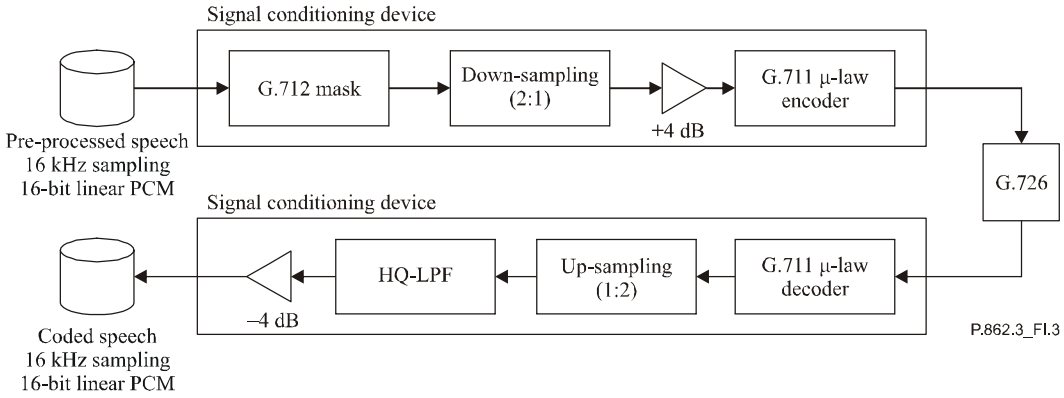


Figure I.3/P.862.3 – Processing of G.726

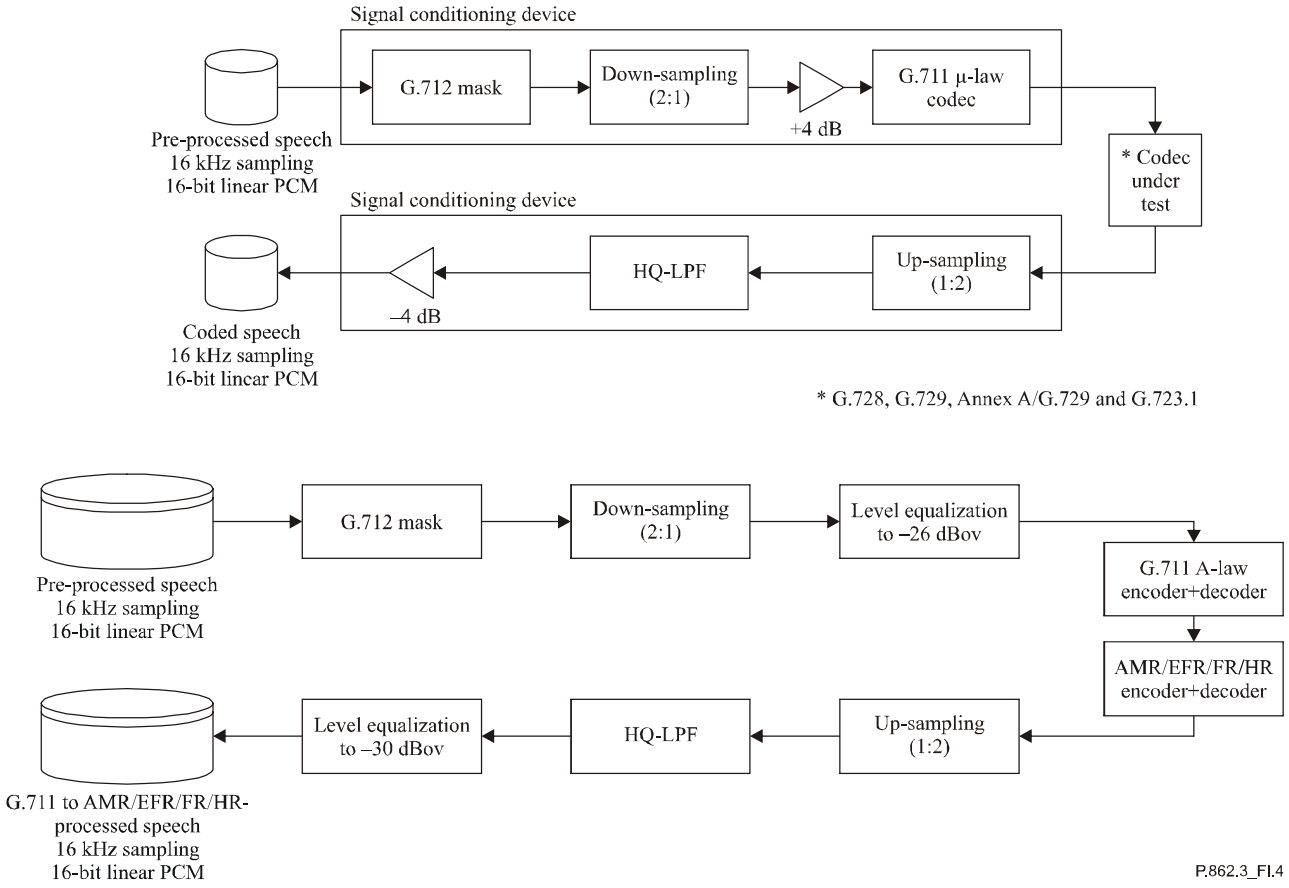


Figure I.4/P.862.3 – Processing of G.728, G.729, Annex A/G.729 G.723.1 and GSM-AMR/EFR/FR/HR

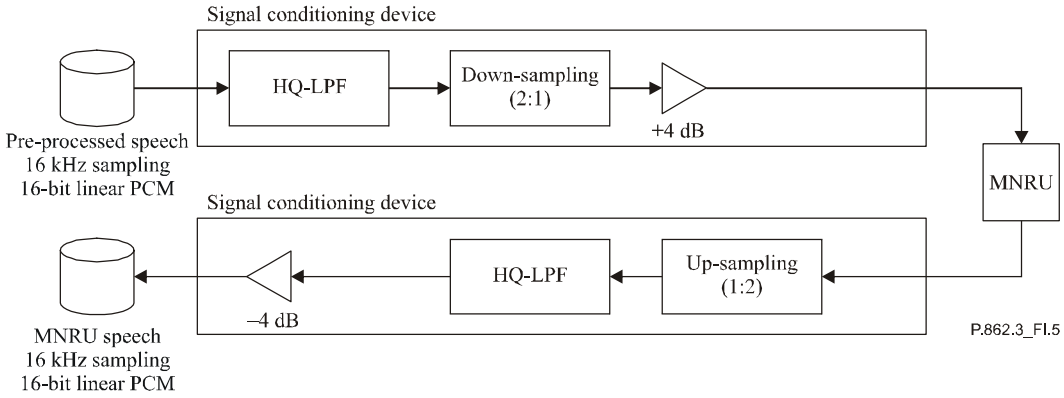


Figure I.5/P.862.3 – Processing of MNRU

The signal processing procedure in Figures I.1 through I.5 is described below.

I.2 Pre-processing of source speech

In the pre-processing part, the source speech goes through a transmission-side modified-IRS filter and its speech level is equalized to –30 dBov as illustrated in Figure I.1. Pre-processed speech files are input signals to a signal conditioning device. Using the software tool provided in ITU-T Rec. G.191, which is called STL2000 hereafter, pre-processed speech can be made by the following commands:

```
$ filter –q mod IRS16 file.inp file.irs
$ sv56demo –q file.irs file.pre 256 1 0 -30
```

I.3 Processing of G.711

Pre-processed speech is masked according to ITU-T Rec. G.712 and down-sampled. Down-sampled speech is equalized in level to –26 dBov and its output becomes the input signal to the G.711 encoder. Processing commands are given below.

```
$ filter -q -down PCM file.pre tmpfile1
$ sv56demo –q tmpfile1 g711.inp 256 1 0 –26 8000
```

ITU-T Rec. G.711 μ-law encoding and decoding processing is carried out using the following STL2000 command.

```
$ g711demo u lili g711.inp g711.dec
```

Decoded speech is up-sampled from 8 to 16 kHz and its level is equalized to –30 dBov. The G.711 coded speech can be obtained by the following STL2000 commands:

```
$ filter -q -up HQ2 g711.dec tmpfile2
$ scaledemo -q -gain 0.63095 tmpfile2 g711.out
```

I.4 Processing of G.726

In G.726 processing, the G.711 μ-law encoder's output becomes the input of the G.726 coder and its output is input to the G.711 μ-law decoder. G.726-coded speech can be obtained by the following STL2000 commands:

```
$ filter -q -down PCM infile tmpfile1
$ sv56demo -q tmpfile1 tmpfile2 256 1 0 -26
$ g711demo u lilo tmpfile2 tmpfile3
$ g726demo u lolo {40|32|24|16} tmpfile3 tmpfile4
$ g711demo u loli tmpfile4 tmpfile5
```

```
$ filter -q -up HQ2 tmpfile5 tmpfile6
$ scaledemo -q -gain 0.63095 tmpfile6 outfile
```

I.5 Processing of G.728, G.729, Annex A/G.729 , and G.723.1

In these codecs, the input signal to the codec under test is the output of a G.711 μ-law codec, i.e., *g711.dec* as mentioned in Figure I.2. When the outputs of the codec under test are denoted by {g728|g729|g729A|G7231}.dec, coded speech is obtained by the following STL2000 commands.

```
$ filter -q -up HQ2 {g728|g729|g729A|g7231}.dec tmpfile2
$ scaledemo -q -gain 0.63095 tmpfile2 {g728|g729|g729A|g7231}.out
```

I.6 Processing of MNRU

Pre-processed speech is down-sampled without changing the frequency response of the input signal, its level is equalized to –26 dBov, and its output becomes the input signal to MNRU. The output of MNRU is up-sampled from 8 to 16 kHz and its level is equalized to –30 dBov. ***Q***-dB MNRU speech can be obtained by the following STL2000 commands.

```
$ filter -q -down HQ2 infile tmpfile1
$ sv56demo -q tmpfile1 tmpfile2 256 1 0 -26 8000
$ mnrudemo tmpfile2 tmpfile3 128 1 0 Q
$ filter -q -up HQ2 tmpfile3 tmpfile4
$ scaldemo -q -gain 0.63095 tmpfile3 mnruQ.out
```

Table I.1/P.862.3 – The P.862.1 reference values for ITU-T standardized codecs

Language	File name	G.711		G.726				G.728	G.729	G.729A	G.723.1	
		μ-law	A-law	16 kbit/s	24 kbit/s	32 kbit/s	40 kbit/s				5.3 kbit/s	6.3 kbit/s
American English	Female 1 (0.00- 7.97 s).wav	4.46	4.28	2.50	3.34	3.89	4.18	3.95	3.95	3.80	3.65	3.81
	Female 2 (0.00- 8.06 s).wav	4.45	4.42	3.12	3.76	4.07	4.33	4.27	4.08	3.99	3.67	3.80
	Male 1 (0.00- 8.44 s).wav	4.49	4.36	2.86	3.82	4.25	4.35	4.19	4.17	4.13	3.90	3.97
	Male 2 (0.00- 7.96 s).wav	4.47	4.31	2.97	3.80	4.21	4.40	4.22	4.15	4.06	3.95	4.07
	Average	**4.47**	**4.34**	**2.86**	**3.69**	**4.11**	**4.32**	**4.16**	**4.09**	**4.00**	**3.80**	**3.92**
Chinese	Female 1 (0.00-10.87 s).wav	4.46	4.48	2.38	3.42	4.21	4.34	3.86	3.72	3.65	3.10	3.33
	Female 1b (0.00-13.39 s).wav	4.42	4.43	2.29	3.34	4.07	4.26	3.98	3.80	3.75	3.26	3.49
	Female 2 (0.00-13.32 s).wav	4.50	4.50	2.26	3.29	4.02	4.37	4.06	3.88	3.75	3.33	3.53
	Female 2b (0.00-13.39 s).wav	4.50	4.50	2.38	3.30	4.03	4.33	4.03	3.87	3.72	3.39	3.59
	Male 1 (0.00-12.15 s).wav	4.44	4.48	2.64	3.58	4.23	4.28	4.19	3.83	3.75	3.36	3.51
	Male 1a (0.00-12.91 s).wav	4.52	4.51	2.84	3.77	4.21	4.37	4.18	4.06	4.00	3.65	3.89
	Male 2 (0.00-12.50 s).wav	4.49	4.48	2.74	3.74	4.30	4.44	4.18	3.99	3.89	3.62	3.78
	Male 2b (0.00-12.82 s).wav	4.50	4.40	2.89	3.90	4.29	4.43	4.16	3.95	3.89	3.35	3.55
	Average	**4.48**	**4.47**	**2.55**	**3.55**	**4.18**	**4.35**	**4.08**	**3.89**	**3.80**	**3.38**	**3.59**
English	Female 1 (0.00- 8.00 s).wav	4.50	4.49	2.69	3.37	3.88	4.21	3.89	3.72	3.58	3.42	3.59
	Female 2 (0.00- 8.00 s).wav	4.48	4.46	2.81	3.44	3.92	4.24	4.00	3.91	3.80	3.60	3.67
	Male 1 (0.00- 8.00 s).wav	4.50	4.45	3.03	3.53	3.96	3.99	4.08	3.88	3.88	3.67	3.81
	Male 2 (0.00- 8.00 s).wav	4.51	4.48	2.94	3.79	4.24	4.34	4.05	3.73	3.60	3.64	3.83
	Average	**4.50**	**4.47**	**2.87**	**3.54**	**4.01**	**4.20**	**4.01**	**3.81**	**3.72**	**3.59**	**3.73**
French	Female 1 (0.00-10.04 s).wav	4.50	4.47	3.06	3.84	4.28	4.42	4.21	3.85	3.77	3.59	3.69
	Female 2 (0.00-10.04 s).wav	4.51	4.48	2.76	3.64	4.15	4.41	4.03	3.77	3.64	3.39	3.56
	Male 1 (0.00-12.18 s).wav	4.50	4.46	3.09	3.79	4.18	4.32	4.09	3.84	3.82	3.45	3.60
	Male 2 (0.00-10.04 s).wav	4.52	4.48	3.33	3.92	4.28	4.40	4.25	4.00	3.91	3.70	3.88
	Average	**4.51**	**4.47**	**3.06**	**3.80**	**4.22**	**4.39**	**4.15**	**3.87**	**3.79**	**3.54**	**3.69**
German	female1 (0.00- 8.00 s).wav	4.49	4.48	2.68	3.46	4.02	4.27	4.04	3.86	3.69	3.54	3.75
	female2 (0.00- 8.00 s).wav	4.48	4.46	2.84	3.65	4.24	4.40	4.13	4.07	3.89	3.61	3.82
	male1 (0.00- 8.00 s).wav	4.50	4.47	2.99	3.72	4.27	4.41	4.09	3.95	3.87	3.56	3.84
	male2 (0.00- 8.00 s).wav	4.50	4.46	2.86	3.47	4.03	4.35	4.12	4.07	4.01	3.75	3.91
	Average	**4.50**	**4.47**	**2.84**	**3.58**	**4.14**	**4.36**	**4.09**	**3.99**	**3.87**	**3.62**	**3.83**
Italian	Female 1 (0.00-20.60 s).wav	4.49	4.40	2.38	3.23	3.81	4.25	3.80	3.75	3.62	3.28	3.49
	Female 2 (0.00-21.78 s).wav	4.48	4.39	2.72	3.68	4.14	4.34	4.16	3.95	3.87	3.57	3.75
	Male 1 (0.00-18.13 s).wav	4.50	4.44	2.61	3.50	4.01	4.33	4.12	3.88	3.81	3.59	3.75
	Male 2 (0.00-20.86 s).wav	4.51	4.43	3.05	3.94	4.28	4.41	4.18	4.12	4.05	3.73	3.95
	Average	**4.49**	**4.41**	**2.69**	**3.60**	**4.07**	**4.33**	**4.07**	**3.93**	**3.84**	**3.55**	**3.74**
Japanese	Female 1 (0.00- 7.60 s).wav	4.46	4.36	2.22	2.97	3.65	4.11	3.76	3.70	3.61	3.25	3.40
	Female 2 (0.00- 7.31 s).wav	4.48	4.38	2.45	3.41	4.12	4.40	3.96	3.82	3.73	3.34	3.59
	Male 1 (0.00- 7.13 s).wav	4.47	4.39	2.36	3.06	3.61	4.20	3.89	3.88	3.74	3.38	3.53
	Male 2 (0.00- 7.45 s).wav	4.49	4.42	2.95	3.90	4.32	4.45	4.31	4.08	4.00	3.83	3.93
	Average	**4.47**	**4.39**	**2.49**	**3.35**	**3.95**	**4.30**	**4.00**	**3.87**	**3.78**	**3.46**	**3.62**
Spanish (US)	Female 1 (0.00- 8.00 s).wav	4.47	4.40	2.33	3.05	3.73	4.17	4.02	3.84	3.71	3.42	3.64
	Female 2 (0.00- 8.00 s).wav	4.40	4.31	2.30	2.92	3.48	4.04	3.77	3.65	3.43	3.07	3.22
	Male 1 (0.00- 8.00 s).wav	4.46	4.30	2.86	3.64	4.19	4.36	4.05	3.83	3.76	3.69	3.77
	Male 2 (0.00- 8.00 s).wav	4.49	4.42	2.76	3.72	4.22	4.40	4.09	3.86	3.85	3.60	3.72
	Average	**4.46**	**4.36**	**2.56**	**3.34**	**3.93**	**4.25**	**3.99**	**3.80**	**3.69**	**3.45**	**3.60**

NOTE – Grey cells indicate samples that do not meet ITU-T Rec. P.501 requirements.

Table I.2/P.862.3 – The P.862.1 reference values for MNRU

Language	File name	MNRU								
		5 dB	10 dB	15 dB	20 dB	25 dB	30 dB	35 dB	40 dB	45 dB
American English	Female 1 (0.00- 7.97 s).wav	1.80	2.39	3.03	3.68	4.15	4.37	4.47	4.50	4.52
	Female 2 (0.00- 8.06 s).wav	2.09	2.67	3.24	3.80	4.16	4.35	4.41	4.44	4.44
	Male 1 (0.00- 8.44 s).wav	1.93	2.58	3.30	4.01	4.32	4.46	4.51	4.52	4.53
	Male 2 (0.00- 7.96 s).wav	1.99	2.65	3.34	3.90	4.23	4.39	4.45	4.47	4.48
	Average	**1.95**	**2.57**	**3.23**	**3.85**	**4.22**	**4.39**	**4.46**	**4.48**	**4.49**
Chinese	Female 1 (0.00-10.87 s).wav	1.41	1.84	2.45	3.19	3.83	4.20	4.40	4.49	4.52
	Female 1b (0.00-13.39 s).wav	1.35	1.74	2.36	3.07	3.66	4.02	4.27	4.36	4.38
	Female 2 (0.00-13.32 s).wav	1.46	1.91	2.54	3.33	3.99	4.33	4.46	4.51	4.54
	Female 2b (0.00-13.39 s).wav	1.52	2.02	2.72	3.51	4.10	4.35	4.42	4.44	4.45
	Male 1 (0.00-12.15 s).wav	1.72	2.29	3.02	3.71	4.16	4.33	4.37	4.38	4.38
	Male 1a (0.00-12.91 s).wav	1.78	2.44	3.22	3.91	4.26	4.43	4.51	4.54	4.54
	Male 2 (0.00-12.50 s).wav	1.67	2.22	3.02	3.78	4.19	4.36	4.46	4.51	4.52
	Male 2b (0.00-12.82 s).wav	1.81	2.46	3.25	3.91	4.29	4.42	4.46	4.47	4.47
	Average	**1.57**	**2.10**	**2.82**	**3.56**	**4.08**	**4.31**	**4.42**	**4.47**	**4.48**
English	Female 1 (0.00- 8.00 s).wav	1.97	2.60	3.27	3.86	4.22	4.41	4.49	4.52	4.53
	Female 2 (0.00- 8.00 s).wav	1.90	2.51	3.16	3.64	4.06	4.34	4.45	4.50	4.52
	Male 1 (0.00- 8.00 s).wav	2.41	3.06	3.65	4.11	4.38	4.47	4.51	4.52	4.53
	Male 2 (0.00- 8.00 s).wav	1.90	2.55	3.32	3.90	4.27	4.45	4.51	4.54	4.54
	Average	**2.03**	**2.68**	**3.35**	**3.89**	**4.24**	**4.42**	**4.49**	**4.52**	**4.53**
French	Female 1 (0.00-10.04 s).wav	1.93	2.55	3.22	3.82	4.23	4.42	4.50	4.53	4.54
	Female 2 (0.00-10.04 s).wav	1.74	2.31	3.02	3.70	4.15	4.40	4.50	4.53	4.54
	Male 1 (0.00-12.18 s).wav	2.07	2.76	3.47	3.96	4.23	4.41	4.49	4.53	4.54
	Male 2 (0.00-10.04 s).wav	2.36	3.14	3.88	4.30	4.46	4.51	4.53	4.54	4.54
	Average	**2.01**	**2.68**	**3.41**	**3.96**	**4.27**	**4.44**	**4.51**	**4.53**	**4.54**
German	female1 (0.00- 8.00 s).wav	1.70	2.28	3.00	3.65	4.19	4.41	4.49	4.53	4.54
	female2 (0.00- 8.00 s).wav	1.74	2.26	2.94	3.64	4.13	4.33	4.47	4.51	4.53
	male1 (0.00- 8.00 s).wav	1.88	2.47	3.17	3.75	4.20	4.42	4.51	4.53	4.54
	male2 (0.00- 8.00 s).wav	2.09	2.71	3.40	3.97	4.33	4.46	4.51	4.53	4.54
	Average	**1.84**	**2.43**	**3.13**	**3.76**	**4.21**	**4.41**	**4.49**	**4.53**	**4.54**
Italian	Female 1 (0.00-20.60 s).wav	1.52	2.01	2.69	3.44	4.01	4.33	4.46	4.51	4.53
	Female 2 (0.00-21.78 s).wav	1.73	2.25	2.92	3.58	4.12	4.38	4.47	4.51	4.53
	Male 1 (0.00-18.13 s).wav	1.79	2.43	3.24	3.94	4.33	4.46	4.51	4.53	4.54
	Male 2 (0.00-20.86 s).wav	2.09	2.86	3.65	4.16	4.40	4.49	4.52	4.53	4.54
	Average	**1.76**	**2.37**	**3.13**	**3.80**	**4.23**	**4.42**	**4.49**	**4.52**	**4.53**
Japanese	Female 1 (0.00- 7.60 s).wav	1.49	1.94	2.57	3.32	3.92	4.27	4.43	4.50	4.52
	Female 2 (0.00- 7.31 s).wav	1.48	1.91	2.51	3.27	3.99	4.35	4.48	4.52	4.53
	Male 1 (0.00- 7.13 s).wav	1.55	2.02	2.65	3.39	4.05	4.36	4.47	4.52	4.53
	Male 2 (0.00- 7.45 s).wav	1.74	2.27	3.01	3.77	4.24	4.45	4.51	4.53	4.54
	Average	**1.56**	**2.03**	**2.68**	**3.44**	**4.05**	**4.36**	**4.47**	**4.52**	**4.53**
Spanish (US)	Female 1 (0.00- 8.00 s).wav	1.49	1.95	2.51	3.16	3.81	4.24	4.43	4.50	4.53
	Female 2 (0.00- 8.00 s).wav	1.61	2.08	2.62	3.19	3.81	4.19	4.40	4.49	4.52
	Male 1 (0.00- 8.00 s).wav	2.02	2.70	3.53	4.14	4.39	4.48	4.51	4.52	4.53
	Male 2 (0.00- 8.00 s).wav	1.65	2.24	3.04	3.73	4.23	4.44	4.51	4.53	4.54
	Average	**1.67**	**2.23**	**2.92**	**3.59**	**4.08**	**4.35**	**4.46**	**4.51**	**4.53**

NOTE – Grey cells indicate samples that do not meet ITU-T Rec. P.501 requirements.

Table I.3/P.862.3 – The P.862 reference values for GSM standardized codecs

Language	File name	AMR								EFR	FR	HR
		12.2 kbit/s	10.2 kbit/s	7.95 kbit/s	7.4 kbit/s	6.7 kbit/s	5.9 kbit/s	5.15 kbit/s	4.75 kbit/s			
American English	Female 1 (0.00- 7.97 s).wav	3.87	3.75	3.60	3.61	3.52	3.43	3.33	3.18	3.94	3.03	3.20
	Female 2 (0.00- 8.06 s).wav	4.13	4.07	3.96	3.92	3.79	3.72	3.62	3.49	4.08	3.70	3.54
	Male 1 (0.00- 8.44 s).wav	4.10	4.03	3.94	3.97	3.88	3.84	3.64	3.50	4.19	3.69	3.46
	Male 2 (0.00- 7.96 s).wav	4.10	4.06	4.03	4.01	3.91	3.85	3.73	3.70	4.20	3.59	3.72
	Average	**4.05**	**3.98**	**3.88**	**3.88**	**3.78**	**3.71**	**3.58**	**3.47**	**4.10**	**3.51**	**3.48**
Chinese	Female 1 (0.00-10.87 s).wav	3.94	3.81	3.46	3.50	3.27	3.15	2.99	3.00	3.98	3.18	2.94
	Female 1b (0.00-13.39 s).wav	3.97	3.82	3.52	3.45	3.36	3.22	3.00	2.95	4.04	3.02	2.95
	Female 2 (0.00-13.32 s).wav	4.08	3.99	3.71	3.62	3.53	3.38	3.21	3.07	4.12	3.07	3.01
	Female 2b (0.00-13.39 s).wav	4.08	4.01	3.74	3.74	3.62	3.40	3.21	3.13	4.10	3.03	3.04
	Male 1 (0.00-12.15 s).wav	3.94	3.84	3.66	3.68	3.54	3.45	3.22	3.22	4.06	3.50	3.26
	Male 1a (0.00-12.91 s).wav	4.23	4.14	3.95	3.92	3.78	3.65	3.34	3.35	4.20	3.65	3.36
	Male 2 (0.00-12.50 s).wav	4.07	3.93	3.78	3.77	3.62	3.46	3.22	3.21	4.15	3.72	3.30
	Male 2b (0.00-12.82 s).wav	4.16	4.15	3.94	3.94	3.80	3.64	3.40	3.27	4.23	3.64	3.44
	Average	**4.06**	**3.96**	**3.72**	**3.70**	**3.57**	**3.42**	**3.20**	**3.15**	**4.11**	**3.35**	**3.16**
English	Female 1 (0.00- 8.00 s).wav	4.00	3.82	3.63	3.62	3.42	3.31	3.21	3.11	3.99	3.27	3.07
	Female 2 (0.00- 8.00 s).wav	3.81	3.78	3.65	3.62	3.56	3.49	3.36	3.34	3.78	3.31	3.28
	Male 1 (0.00- 8.00 s).wav	4.01	3.88	3.75	3.67	3.46	3.57	3.23	3.01	4.03	3.54	3.37
	Male 2 (0.00- 8.00 s).wav	4.06	3.83	3.73	3.68	3.53	3.48	3.18	2.98	4.10	3.75	3.49
	Average	**3.97**	**3.83**	**3.69**	**3.65**	**3.49**	**3.46**	**3.25**	**3.11**	**3.98**	**3.47**	**3.30**
French	Female 1 (0.00-10.04 s).wav	4.11	4.01	3.79	3.83	3.66	3.42	3.33	3.30	4.04	3.49	3.37
	Female 2 (0.00-10.04 s).wav	3.91	3.85	3.57	3.54	3.40	3.31	3.22	2.98	3.83	3.19	3.24
	Male 1 (0.00-12.18 s).wav	4.00	3.88	3.71	3.72	3.54	3.37	3.23	3.14	4.07	3.49	3.32
	Male 2 (0.00-10.04 s).wav	4.11	4.05	3.85	3.91	3.75	3.57	3.37	3.28	4.17	3.84	3.30
	Average	**4.03**	**3.95**	**3.73**	**3.75**	**3.59**	**3.42**	**3.29**	**3.18**	**4.03**	**3.50**	**3.31**
German	female1 (0.00- 8.00 s).wav	4.08	3.98	3.65	3.60	3.54	3.36	3.15	3.06	4.06	3.40	3.20
	female2 (0.00- 8.00 s).wav	4.21	4.14	3.93	3.88	3.76	3.63	3.53	3.47	4.17	3.54	3.37
	male1 (0.00- 8.00 s).wav	4.12	4.08	3.90	3.88	3.78	3.66	3.55	3.50	4.19	3.82	3.43
	male2 (0.00- 8.00 s).wav	4.17	4.07	3.97	3.92	3.77	3.69	3.65	3.56	4.21	3.70	3.38
	Average	**4.14**	**4.07**	**3.86**	**3.82**	**3.72**	**3.58**	**3.47**	**3.40**	**4.16**	**3.62**	**3.35**
Italian	Female 1 (0.00-20.60 s).wav	3.80	3.67	3.51	3.41	3.34	3.24	3.12	2.92	3.81	2.83	3.00
	Female 2 (0.00-21.78 s).wav	4.09	4.04	3.88	3.86	3.74	3.61	3.40	3.27	4.14	3.29	3.32
	Male 1 (0.00-18.13 s).wav	4.03	3.95	3.82	3.78	3.64	3.49	3.32	3.20	4.13	3.37	3.17
	Male 2 (0.00-20.86 s).wav	4.23	4.15	4.00	4.07	3.89	3.84	3.60	3.47	4.27	3.58	3.47
	Average	**4.04**	**3.95**	**3.80**	**3.78**	**3.65**	**3.55**	**3.36**	**3.22**	**4.08**	**3.27**	**3.24**
Japanese	Female 1 (0.00- 7.60 s).wav	3.89	3.75	3.53	3.41	3.40	3.28	3.08	3.10	3.87	2.92	2.92
	Female 2 (0.00- 7.31 s).wav	3.92	3.85	3.54	3.62	3.42	3.27	3.20	3.08	3.99	2.82	3.09
	Male 1 (0.00- 7.13 s).wav	3.87	3.81	3.59	3.50	3.42	3.28	3.19	3.12	3.91	2.89	2.91
	Male 2 (0.00- 7.45 s).wav	4.18	4.08	3.92	3.94	3.84	3.73	3.59	3.44	4.21	3.63	3.55
	Average	**3.97**	**3.87**	**3.64**	**3.61**	**3.52**	**3.39**	**3.27**	**3.19**	**4.00**	**3.06**	**3.12**
Spanish (US)	Female 1 (0.00- 8.00 s).wav	4.03	3.96	3.63	3.68	3.50	3.30	3.23	3.16	3.96	3.03	3.20
	Female 2 (0.00- 8.00 s).wav	3.63	3.48	3.23	3.26	3.05	3.03	2.88	2.80	3.73	2.68	2.72
	Male 1 (0.00- 8.00 s).wav	3.98	3.61	3.67	3.49	3.51	3.38	3.23	3.15	4.15	3.44	3.19
	Male 2 (0.00- 8.00 s).wav	4.04	3.87	3.71	3.58	3.52	3.28	3.11	3.14	4.09	3.48	3.36
	Average	**3.92**	**3.73**	**3.56**	**3.50**	**3.40**	**3.25**	**3.11**	**3.06**	**3.99**	**3.16**	**3.12**

NOTE – Grey cells indicate samples that do not meet ITU-T Rec. P.501 requirements.

Appendix II

Test databases for P.862/P.862.1

The test databases comprised network conditions which corresponded to fixed, wireless and VoIP applications. A large set of codecs used by different technologies (such as GSM-FR, GSM-EFR, GSM-AMR, CDMA-EVRC, IS136-ACELP, G.711, G.726, G.728, G.729, JDC-HR) has been included in the P.862 and P.862.1 tests and analysis.

Table II.1 summarizes the content of all the databases on which PESQ (P.862 and P.862.1) have been validated.

Table II.1/P.862.3 – Summary of the databases' type and content

Description	Conditions
8 kbit/s characterization interworking with standards, P-suppl. 23, exp 1	enc/dec + trans.; codecs: G.711, G.726, G.728, G.729, Is-54, GSM-FR, JDC-HR.
8 kbit/s characterization channel errors and noise, P-suppl. 23, exp 3	enc/dec + trans + ErorrPatterns + BGN, codecs: G.729.
Live wireless networks	– Wireless networks: IS-136, CDMA, iDEN, AMPS; GSM-US, GSM-Europe; – Codecs: IS-54, 8 kbit/s ACELP, 13 kbit/s QCELP, GSM-FR and EFR, CDMA- EVRC.
Codecs, errors patterns, transcodings, noise	– enc/dec + ErrorPatterns + BGN; codecs: G.711, G.726, G.728, G.729, GSM-FR; – enc/dec + ErrorPatterns (C/I levels) + BGN; codecs: G.711, G.723 + transcodings, 8 kbit/s ACELP, EVRC, GSM-EFR and FR; – enc/dec + trans + ErrorPatterns ("bad" & "good" cond) + BGN, ATM/ISDN/POTS; codecs: G.729, G.728, GSM-FR, GSM-HR.
Background noise test, GSM and fixed networks	– enc/dec + VAD + VQE (NR); codecs: GSM-FR, G.729; – enc/dec + ErrorPatterns; codecs: GSM-FR, GSM-AMR-HR.
DTX, frame/burst erasure and VAD test, AMR, GSM and fixed networks	enc/dec + ErrorPatterns; codecs: GSM-FR, G.726, G.728, G.729, AMR.
AMR + error patterns on hopping and non-hopping GSM channels	enc/dec + ErrorPatterns; codecs: AMR475, AMR590, AMR740, AMR122.
AMR + error patterns with unequal and equal error protection	enc/dec + ErrorPatterns; codecs: AMR515, AMR740, AMR102, AMR122, AMR475, AMR590, AMR670, AMR795.
VoIP error pattern and packet loss; noise	enc/dec + trans + ErrorPatterns + PacketLoss + BGN.
VoIP packet loss scenarios test	VoIP: PktInsert, PktDelete, PktMute.
VoIP codecs	VoIP: G.723.1, G.728, G.711, G.729, G.726, PDC-HR.
VoIP conditions	PLC and jitter buffer managm. (G.711, G.729, G.723 at 6.3 and 5.3 kbit/s), variable packet length.
Field VoIP conditions	VoIP: Internet PABX, IP Gateways.

Appendix III

Report of P.862/P.862.1 measurements

The P.862/P.862.1 measurements should be reported based on the algorithm's accuracy presented in 10.3 (Tables 1 and 2).

III.1 Report and interpretation of the average PESQ results

As mentioned in 10.1, it is recommended in the case of controlled simulated network test conditions to perform averages per-condition, using at least four talkers.

The averaged P.862.1 scores are expected to be determined with a prediction error of maximum ±95% confidence interval upper limit.

In the case of field network testing, there are situations when it is desirable to roughly estimate the speech quality provided by the network within a certain test area, and/or during a time window. In this case, P.862.1 scores are averaged along a test route and/or a time window. The averaged P.862.1 scores are expected to be determined with a prediction error of maximum ±95% confidence interval upper limit. Therefore, assuming for example that the tested network is a CDMA network then it is expected that the average prediction error of the averaged P.862.1 measurements is less than or equal to 0.462 MOS.

It should be noted though that for both cases presented above, simulated and field conditions, there is a risk of 5% that the measurement error is higher than the 95% upper limit of the prediction error. In addition, as mentioned in 10.2, the measurement scenarios need to be similar to the ones presented in Table 1.

III.2 Report and interpretation of individual PESQ measurements' results

As mentioned in 10.1, in the case of live field network testing a per-sample quality evaluation, due to uncontrolled time-varying transmission channels, is required.

There are two recommended reporting procedures, depending on the type of the PESQ measurements, which are used for the network's speech quality analysis.

The first type is the averaged P.862.1 score, which is discussed in III.1 and provides a rough estimate of the speech quality within a certain area or during a time window.

The other type of score is the individual P.862.1 measurement when troubleshooting of the network is required.

In the case of individual measurements, it is recommended to calculate a histogram of the P.862.1 measurements along the MOS scale. As an analysis procedure in this case, it is recommended to consider a subjective speech quality threshold and to impose a minimum required percentage of scores lying above this subjective threshold or equivalently a minimum required probability with which the P.862.1 measurements in the tested network lie above the subjective threshold. In this way, the speech quality based on individual P.862.1 scores is represented by the probability or percentage of scores that lie above the imposed subjective speech quality threshold.

For the evaluation of this probability or percentage, two types of errors affect the results. The first error is caused by the calculation of the probability or the percentage. The other error is determined by the P.862/P.862.1 residual error distribution. Due to the P.862.1 measurements' error, not all the points determined to be above (and/or below) the imposed MOS threshold are in reality there. As mentioned in 10.4, the residual error distribution (see Table 2) shows that lower absolute errors are more likely than higher absolute errors. Therefore, when a P.862.1 score is on one side of the MOS threshold based on the actual measurement, the likelihood for it to appear on the other side of this

threshold is higher if the actual score is closer to the threshold. For example, let us consider a MOS threshold of MOS=3 and a CDMA test network. P.862.1 scores between 3 and 3.1 have a probability of 40.44% to fall below the threshold. P.862.1 scores between 3.1 and 3.2 have a probability of only 30.04% to fall below the threshold. Similarly, scores between 2.9 and 3 have a probability of 40.44% and scores between 2.8 and 2.9 have probability of only 30.04% to fall above the 3 MOS threshold.

A recommended method for calculating the P.862.1 measurement error when using individual scores now follows:

The P.862.1 measurement error, due to the error of evaluation of the probability that P.862.1 scores are higher than the imposed MOS threshold, is defined as the standard deviation of a binomial distribution with the probability of occurrence:

$$p = \frac{n}{N}$$

where n represents the number of scores above the threshold and N represents the total number of P.862.1 measurements. The error caused by the probability evaluation Error1 is given therefore by:

$$Error1 = \pm\sqrt{\frac{p \times (1-p)}{N}}$$

and it is characterized by the 95% confidence interval:

$$\pm z\alpha \times Error1$$

where:

z_α = 2 represents the Gaussian quantile for 95% probability (the normal distribution describes quite well the PESQ individual scores' behaviour).

The P.862/P.862.1 measurement error due to the residual error of the PESQ algorithm Error2 is recommended to be calculated as the standard deviation of the binomial distribution that describes the residual error presented in Table 2.

When individual P.862/P.862.1 scores are used, it is therefore recommended to calculate the total measurement error of the speech quality as the square root of the sum of the two above-mentioned squared errors, Error1 and Error2.

Appendix IV

Calibration method for proprietary interfaces

NOTE – The method described here is intended for situations where the exact, required signal levels are unknown and the methods recommended in clause 9 cannot be applied. It cannot be expected to achieve maximum accuracy by this method. It is also important that all level adjustments are either performed in the analogue domain or using a sufficient word length in the digital domain (at least 24 -bits).

IV.1 Calibration of the transmit level (near end) of the test equipment

Adjust the output level of the test equipment to a level which falls approximately in the middle of the operating range of any potential AGC. This can be done by measuring the signal level at the far end terminal and tuning the attenuation of the signal level at the near end to the middle of the range at which the level at the far end remains constant. Alternatively a phone may be used on the far end and the level at the near end is adjusted to the middle of the range which gives a comfortable listening level on the far end. The second method is also preferable in situations where no AGC is present.

IV.2 Calibration of the receive level (far end) of the test equipment

The level at the far end should be adjusted in a way that the attenuation between the recorded file and the reference file is close to 0 dB.

BIBLIOGRAPHY

[B.1] Objective quality evaluation based on ITU-T Rec. P.862 by using long reference speech (NTT), COM12-D008, Jan. 2005.

[B.2] Objective quality measurement using artificial voice signals (NTT), COM12-D145, Sep. 2003.

[B.3] Addition of noise floor to reference speech used in ITU-T Rec. P.862, COM12-D011, Jan. 2005.

[B.4] ANSI/TIA-127-A-2004 (2004), *Enhanced Variable Rate Codec Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*.

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	**Telephone transmission quality, telephone installations, local line networks**
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, 2006