Telephone Speech Quality Standards

for

IP Phone Terminals (handsets)

Measuring method

CES-Q004M-1

March 31, 2008

Communications and Information network Association of Japan
(CIAJ)

Table of Contents

1. Introduction

CIAJ adopted "*Telephone Speech Quality Standards for Wideband IP Phone Terminals (handsets):* CES-Q004-3" on April 1, 2007. This standard, CES-Q004-1, describes the measurement methods for verifying performance and should be used in conjunction with CES-Q004-3.

2. Scope

At present the following two types telephone terminals exist for VoIP communication:

- "*IP-Phones,*" which can be directly connected to IP networks
- Analog telephone sets, which can be connected to IP networks using an adapter (or converter). Such converters are called Terminal Adapters (TA) or Residential Gateways (GW).

This standard describes the method of measuring the performance of IP phones.

For telephone sets without handsets, such as hands-free phones, the measurement method is still under study and is not within the scope of this standard.

This standard refers to wideband terminals compliant to ITU-T Rec. P. 311 in the bandwidth of 150 Hz to 7 KHz and does not include 100 Hz band terminals included in codec standards.

3. References

[1] *CIAJ Standard Telephone Speech Quality Standards for IP Phone Terminals (handsets)* CES-Q003-2

[2] *CIAJ Standard Telephone Speech Quality Standards for Wideband IP Phone Terminals (handsets)* CES-Q004-1

[3] ITU-T Rec. G. 107
The E-model, a computational model for use in transmission planning

[4] ITU-T Rec. G. 122
Influence of national systems on stability and talker echo in international connections

[5] ITU-T Rec. G. 711
Pulse code modulation (PCM) of voice frequencies

[6] ITU-T Rec. G. 311
Transmission characteristics for wideband (150 – 7000 Hz) digital handset telephones

[7] ITU-T Rec. P. 50
Artificial voices

[8] ITU-T Rec. P. 51
Artificial mouth

[9] ITU-T Rec. P. 53
Psophometer for use on telephone-type circuits
[8] ITU-T Rec. P. 57
Artificial ears
[9] ITU-T Rec. P. 64
Determination of sensitivity/frequency characteristics of local telephone systems
[10] ITU-T Rec. P. 79
Calculation of loudness ratings for telephone sets
[11] ITU-T Rec. P. 310
Transmission characteristics for telephone band (300-3400 Hz) digital telephones
[12] ITU-T Rec. P. 800
Methods for subjective determination of transmission quality
[13] ITU-T Rec. P. 833
Methodology for derivation of equipment impairment factors from subjective listening-only tests
[14] ITU-T Rec. P. 834
Methodology for derivation of equipment impairment factors from instrumental models
[15] ITU-T Rec. P. 862
Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs
[16] ITU-T Rec. P. 862.1
Mapping function for transforming P.862 raw result scores to MOS-LQO
[17] ITU-T Rec. P. 1010
Fundamental voice transmission objectives for VoIP terminals and gateways
[18] ITU-T Rec. Y. 1541
Network performance objectives for IP-based services

4. Interfaces

Either the *Direct Approach* or the *Reference Codec Approach* defined in ITU-T Rec. P. 311 *Electrical Interface Specifications* shall be applied.

(1) Functions and configuration

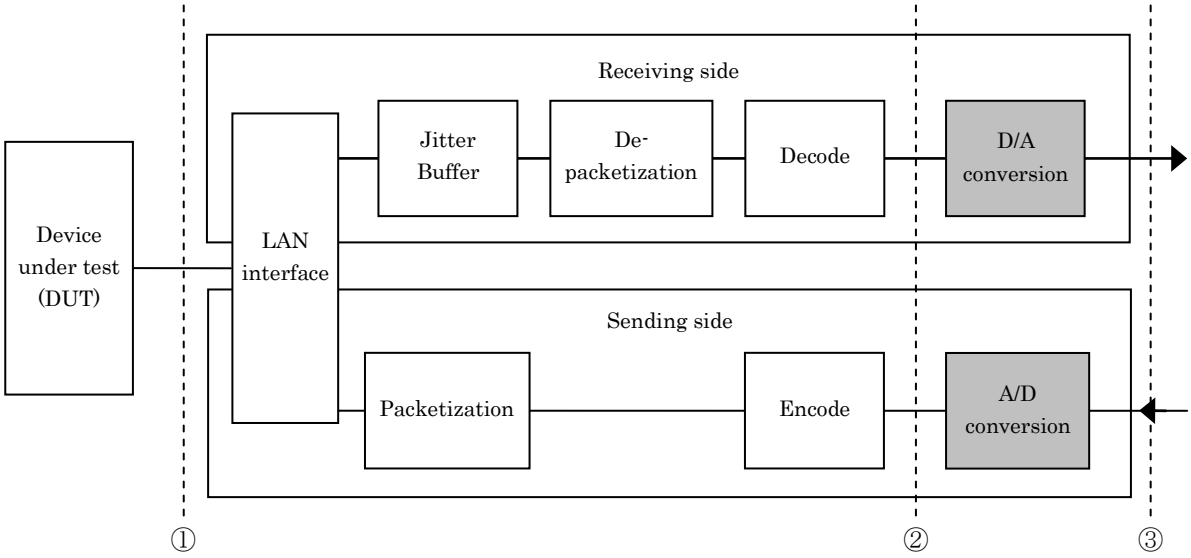


Figure 1/CES-Q004M Configuration of interface unit for connection to DUT (the section from ① to ② or ① to ③)

Note: Neither A/D nor D/A conversion is mandatory.
The external measuring equipment is connected to either digital interface ② or analog interface ③.
The facilities for call set-up are omitted in Figure 1.

The measuring device presumes that the encoding side can transmit according to G. 722 Mode 1. Even if any other encoding format is used as a standard, it shall be possible to switch to G. 722 Mode 1 in some way.
If another encoding format is used, the signal level overload point shall be the same as G. 722 encoding format, or +9 dBmO.

(2) Characteristics at each side[i]

[Sending side]

Propagation delay jitter: less than ±10ms.

Frequency characteristics: The frequency characteristics of the section from ③ to ①, or from ② to ① should meet the requirements of Figure 2, which are the same as ITU-T Rec. P. 311 Fig. B3/P.311.

[i] The absolute propagation delay at the interface unit is not critical. (If the delay in the interface itself is clearly identified, the delay in the DUT can be determined by subtracting this known additional delay from the overall delay.)

[Receiving side]

Propagation delay jitter: less than ±20 ms.

Frequency characteristics: The frequency characteristics of the section from ① to ③, or from ① to ② should meet the requirements of Figure 2, which are the same as ITU-T Fig. 2/P. 311.

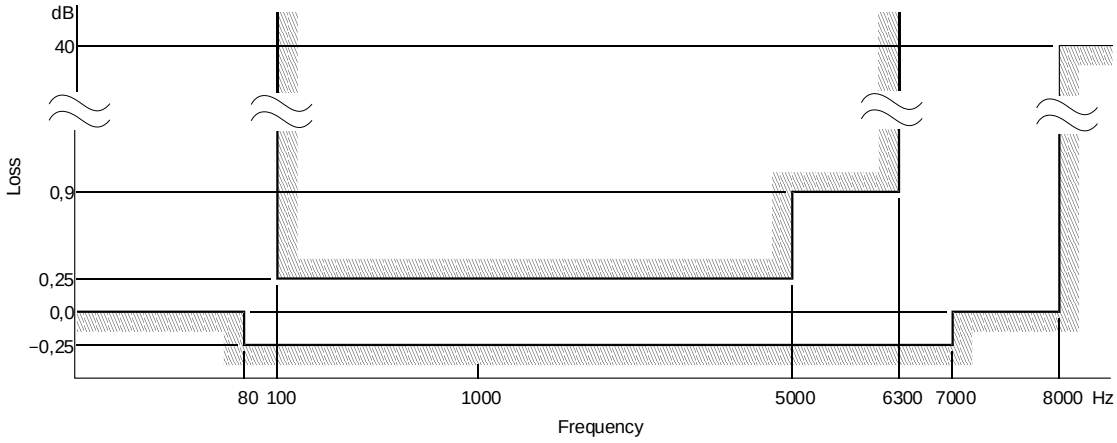


Figure 2/CES-Q004M　Frequency response of the receiving side

i) Distortion

Distortion due to speech coding error or sending-side packet stream jitter can be controlled using existing technology to a level which will not impact measurement results. On the other hand, the receiving side of the interface is more important than the sending side since the influence of jitter contained in the packet stream sent from the DUT is controlled by the management of jitter buffer in the interface unit. Distortion due to speech coding or packet loss can be evaluated using wideband PESQ values derived in accordance with ITU-T Rec. P. 862.

Therefore, only the requirements for distortion on the receiving side are specified.

When Wideband PESQ is used for evaluation, the signal at ① is considered as the input and the signal at either ② or ③ is considered as the output.

The Wideband PESQ value should lie within the area shown in Fig. 3 when external jitter is applied to the section between the test terminal and the interface, using a network simulator.

Note: For this validation, a DUT with negligible jitter should be used.

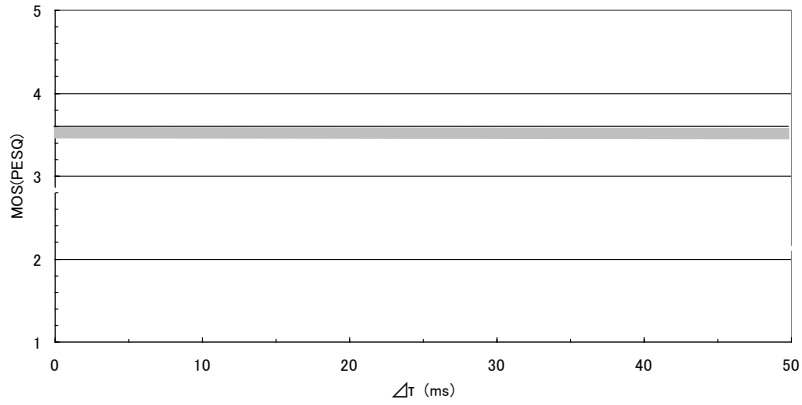


Figure 3/CES-Q004M Permissible minimum value of PESQ against network jitter Δτ for the receiving side of the interface unit

ii) Level accuracy

If coding based on ITU-T Rec. G. 722 is used, digital codes and analog signal levels (dBm) should be related (converted) in accordance with the relationship shown in Note 1 of Table 1a - 2b of G. 711.

If other types of coding law are used, the overload point (maximum fluctuation) should be the same as that of G. 722 coding.

The accuracy of analog signal levels should be within ±0.3dB.[ii]

5. Test signals, Artificial Mouth, Artificial Ear

5.1. Test signals

Either a modified artificial voice generated from original signals conforming to ITU-T Rec. P. 50 (Option 1) or a recorded real speech signal longer than 8 seconds in accordance with P. 800 (Option 2) should be used.

5.1.1. Artificial signal (Option 1)

Rec. P. 50 defines two kinds of artificial voice, representing the spectral characteristics of male and female speech respectively (source signals). The duration of each signal is 10 seconds.

The test signal is generated according to the following procedure.

- The source signal is divided into an even number (at least 4) of short

[ii] Complies to scope of overload level designated in ITU-T Rec. G. 722.

segments (Seg. n in Figure 4) at points where the short-term power envelope falls close to zero, avoiding the middle of individual phonemes. Each segment is approximately 2,500ms long with ±400ms tolerance.

- Even numbered segments of the male voice are replaced by corresponding segments of the female voice. (Alternatively, even numbered segments of the female voice may be replaced by corresponding segments of the male voice.)
- A 700 ms (±150ms) long silent period is inserted between segments.

ITU-T Rec. P. 50 defines artificial voice frequency level as 100 – 8000 Hz.

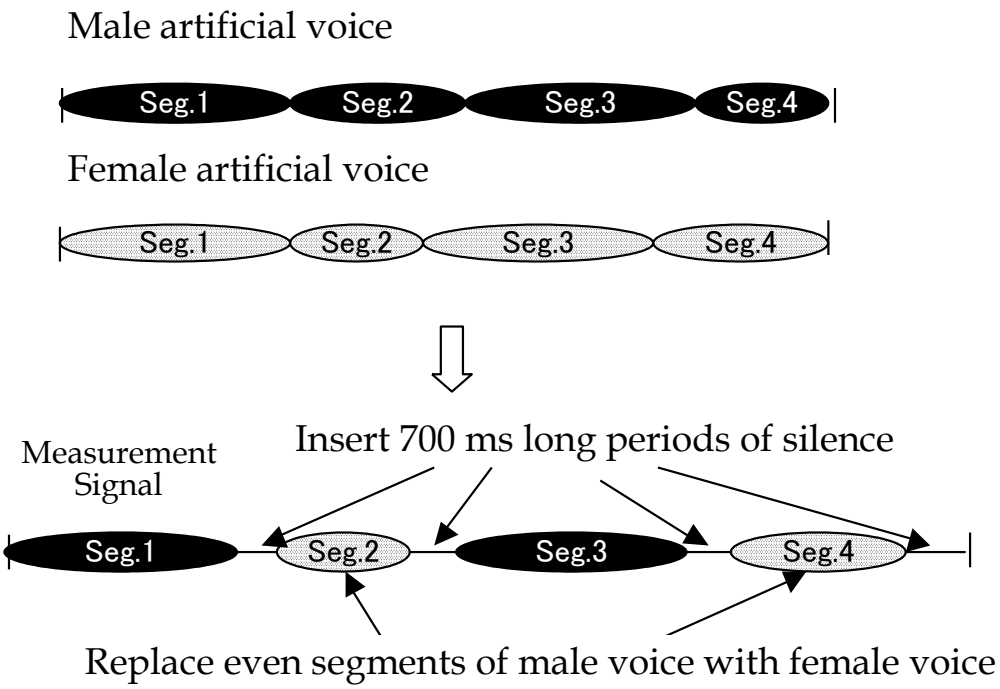


Figure 4/CES-Q003M How to generate the test signal

5.1.2. Real voices (Option 2)

When real voices are used, voice samples spoken by at least two different males and two different females should be used. To measure the sensitivity/frequency response (S.F.R) to derive the loudness rating, WB-PESQ, the average of the measured values obtained for these multiple test signals should be used as the final measured value.

The frequency band for real voices (option 2) shall be in the range of 100 Hz to 7 kHz.

5.2. Artificial mouth

An artificial mouth complying with ITU-T Rec. P. 51 shall be used.

Artificial voice spectrum at MRP point shall be within the range specified in ITU-T Rec. P. 50.

5.3. Artificial ear

Of the artificial ears complying with ITU-T Rec. P. 57, type 3.2 *low leak version* shall be used. The frequency level for ITU-T P. 57 Type 3.2 artificial ear is defined as in the range of 100 – 8000 Hz.

5.3.1. ERP-DRP conversion

The artificial ear measures DRP sound pressure. Therefore, if the loudness rating or WB-PESQ is being calculated, the figures shall be converted to ERP before being applied. The conversion coefficient in ITU-T Rec. P. 57, Table 2a/P. 57 shall be used.

5.3.2. Notes concerning leakage

When using Type 3.2 artificial ear with low leak in compliance with ITU-T Rec. P. 57, the following shall apply.

> No loss correction shall be made in the calculation of loudness ratings.

> Be attentive to surrounding noise when using acoustically open artificial ears, which have an open back side, for surrounding noise may penetrate into the artificial ear.

> When measuring sidetone, sound from the artificial mouth leaking in from the back side of the artificial ear will be mistakenly measured as sidetone. Therefore, the following procedures shall be followed for sidetone:

① Measurement shall be taken using Type 1 artificial ear with no leakage, or Type 3.2 artificial ear designated in P. 57 with no leakage.

② Receiving characteristics shall be determined by the difference in measurements of an artificial ear with no leakage to a Type 3.2 artificial ear designated in P. 57 with leakage.

③ The impact of an artificial ear on sidetone characteristics is the same as receiving side, so the figure attained in ① shall be corrected by the difference calculated in ②.

④ In addition, make DRP-ERP conversions.

Furthermore, DRP-ERP conversions shall be done as usual.

Since complete acoustic isolation at the ear piece is not possible even when using Type 1 or Type 3.2 artificial ears, there are limitations to the measurements when there is a large amount of sidetone attenuation.

6. Loudness ratings

This measurement is performed without any network disturbance, i.e., condition "0." The signals described in Section 5 should be used. The SFR is measured using the same procedure as for digital telephone sets, as defined in ITU-T Rec. P. 64.

For the sending measurement, the average sound pressure level at the mouth reference point (MRP) should be adjusted to -4.7 dBPa* (±3 dB).

For wideband voice encoding format, the generally used ITU-T Rec. G. 722 series overload point of + 9 dBmO and standard signal level of – 15 dBmO shall be applied.

In the extension of ITU-T Rec. 711 for narrowband PCM of voice frequencies, the overload point of + 3.17 dBmO and standard signal level of – 15 dBmO shall apply when using the same overload point as G. 711.

The loudness rating shall be calculated using the SFR and the algorithm defined in ITU-T Rec. P. 79.

7. Delay (Latency)

The signals described in Section 5 should be used.

The delay value is derived by evaluating the short-term cross-correlation between input and output signals. The time alignment process in the PESQ algorithm (in ITU-T Rec. P. 862) can be applied to this procedure.

In measuring the sending delay, the signal at MRP is taken as the input signal, and the signal at either ② or ③ in the receiving part of the interface unit as the output signal. The measured overall delay includes not only the required value of the DUT but also the additional interface unit delay. Subtracting the additional delay from the overall delay gives the actual delay of the DUT.

This operation should be repeated until the total duration becomes longer than 3 minutes. The measured value should be determined as the average over the entire length of the signal before starting the measurement.

The internal state of the test object should be initialized. (If no visible reset button is provided on its housing, turning the power off-and-on can be used as the

* ITU-T Rec. P.50

initialization operation.)

7.1. Sending delay

This measurement should be performed without any network disturbance.

The acoustic signal at the MRP is recorded as the input signal (the handset is not placed in front of the artificial mouth conforming ITU-T Rec. P.51 mentioned in 5) using a half-inch condenser microphone placed in front of artificial mouth. The handset should be removed while this recording is made. In the definition, the difference in delay time due to the positional difference between the MRP and the actual mouthpiece position is ignored.

The output signals should be recorded at either point 2 of Fig. 1 (digitally) or point 3 of Fig. 1 (analog) of the interface unit.

Note) The delay from a single point in the input signal waveform to the corresponding point in an output packet may be inaccurate since the packet output from the DUT is subject to a greater or lesser degree of accompanying jitter even when no network disturbance is present. Averaging over a number of points and packets can increase the reliability of the measurement.

7.2. Receiving delay

Network disturbance effects are provided in accordance with CES-Q004-1.

The load conditions defined in CES-Q004-1 are designated in table 7.1.

Table 7.1.

Element	**IP network performance objectives**
Average delay time, T (ms)	70
Maximum delay time, Ta (ms)	67 . 10
Maximum delay variation, ΔT: Δt_{max} (ms)	20
Average delay variation: Δt_{ave} (ms)	2 . 90
Packet loss ratio (%): Ppl	0 . 1

Delay variation: The change in the instantaneous delay time of the network from Ta to Ta+ΔT.

Probability of ΔT: It is assume to follow an exponential distribution. If ΔT is greater than the packet

transmission interval, packets may arrive out of sequence. The probability of the delay variance ΔT occurring up to and including maximum value Δt_{max} shall be 99.9%.

●Probability of Ppl: It is assumed to follow a uniform distribution (random loss). Busty loss is left as an item for further study.

Refer to Section 11 for probability distribution of delay at 70 ms when exponential distribution delay variation is set at 20 ms.

In other words, if the delay is stated as 70 ms, it is the sum of the minimum delay time (delay time with no variation) and the average value of delay variation. This is because the "*delay time*" measured here is the "*average value*," and the "*variation*" is the difference from this value.

The input signals should enter the encoder of the interface unit either at point 2 of Fig. 1 (digital) or point 3 of Fig. 1 (analog).

The output signals should be those at the ERP (Ear Reference Point), or the entrance to an artificial ear of ITU-T Rec. P. 57, Type 3.2 coupled to the earpiece. In the definition, the propagation time in the air from the earpiece of the DUT to the microphone of the artificial ear, and the electro-acoustic conversion time in the microphone are ignored.

Note) The same precaution should be taken as for the sending delay measurement. Since there is jitter in the packets entering the DUT and, in general, there is no mechanism available within the DUT to absorb the variation, the difference between input and output signals at a single point in time should not be used as the delay time.

8. Echo aspects

Terminal coupling loss is the path loss including acoustic reflection from a receiver to its own transmitter. In terminals such as IP-phones, the echo return loss is determined only by the terminal acoustic coupling loss because they are not subject to any electrical reflection in an item such as a hybrid circuit.

TELR is the sum of TCLw and SLR and RLR.

8.1. Measurement of TCLw

Echo return loss should be measured in accordance with ITU-T Rec. G. 122 Annex A. However, the test signals described in Section 3 of this document should be used. From the frequency aspect, the 1/f weight specified in G. 122 Annex B.4

should be applied to obtain the loss.

While keeping the handset in an anechoic space as described in ITU-T Rec. P.310 Fig.B8/P. 310, the loss is defined as the attenuation from the input of the receiving side of the DUT to the output of sending side. When using the measurement interface unit shown in Fig. 1, the input point is either point 2 (digital) or point 3 (analog) of the sending side and the output point is either point 2 (digital) or point 3 (analog) of the receiving side.

Note) Rules for narrowband exist, but since there are none pertaining to wideband, close values in the Recommendation shall be used.

9. Idle noise

9.1. Sending side: Nc

Psophometrically weighted signal power level of idle noise contained in the payload of packets sent from the DUT output is measured in a quiet environment (ambient noise less than 30 dBA). "A" weighting is specified in ITU-T Rec. P. 53.

9.2. Receiving side: Nfor

While digital codes of signal level 0 (μ-law) of G. 711 from the IP network are input to the DUT, the "A" weighted sound pressure level (dBPa(A)) of the idle noise detected by an ITU-T P. 57 Type 1 artificial ear coupled to the earpiece should be measured. The "A" weighting is shown in standard IEC 60651.

10. Overall audio quality

In past telephone bandwidths, the "*R value*" was used as the indicator for overall audio quality, but nothing is specified by ITU-T for wideband voice communications. Therefore, ITU-T Rec. P. 86 2.2 Wideband PESQ (WB-PESQ) shall be used to calculate MOS for wideband.

10.1. Measurement method

The WB-PESQ designated in P. 862.2 shall be measured.

The sending and receiving sides shall be measured under end-to-end acoustic input output measurement conditions.

Input and output signals for applying PESQ are defined as follows.

(1) Sending side:

Input signals: Output signals from a microphone placed at the defined MRP, in front of an artificial mouth.

Output signals: Digital or analog signals decoded at the interface unit (Point 2 or 3 of the receiving side in Fig.1)[1]

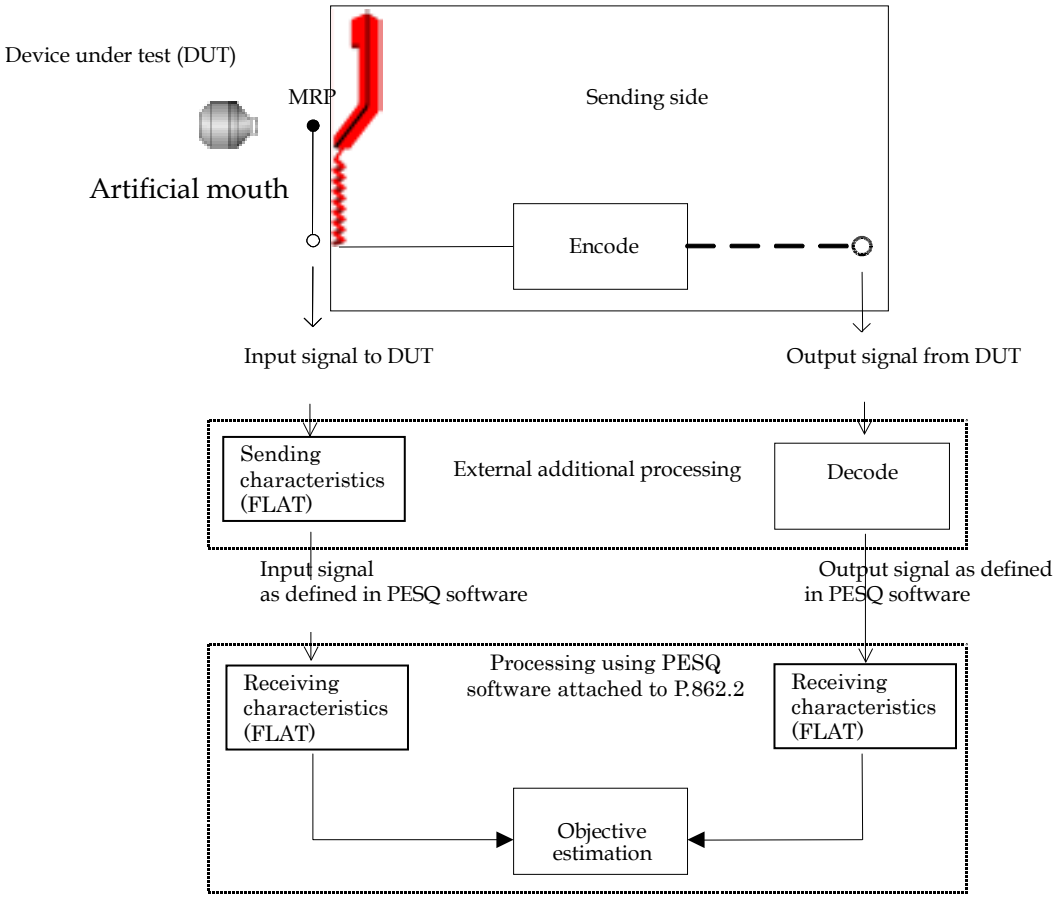


Figure 5/CES-Q003M-1 Definition of input and output points and process flow for deriving WB-PESQ.

(Bold dotted lines show the IP section.)

[1] In the software attached to ITU-T Rec. 862.2, flat receiver characteristics is programmed into the software and applied to the input/output signals.

(2) Receiving side

Input signals: Digital signals input to the encoder at the interface.

Output signals: Output signals from an artificial ear coupled to the earpiece.

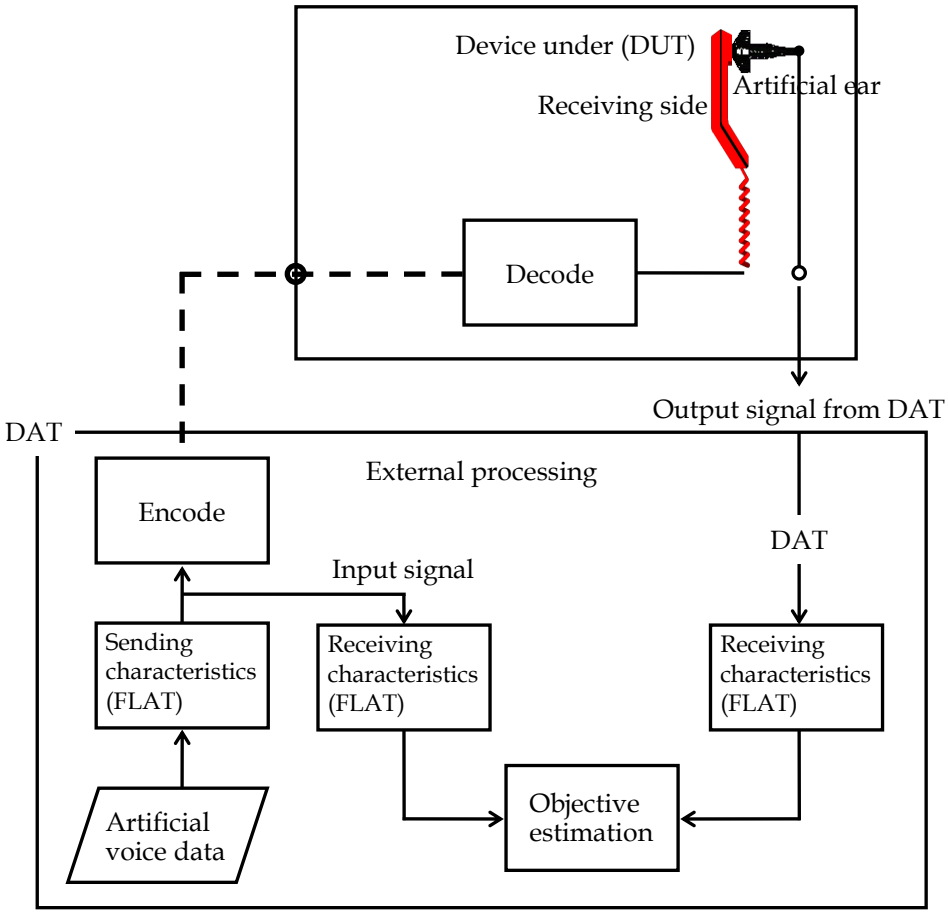


Figure 6/CES-Q003M-1 Input and output signals points and process flow diagram for deriving receiving MOS by applying PESQ

(Bold dotted lines show the IP section.)

(3) Output mapping

The expanded wideband covered in ITU-T Rec. P. 862 includes the following mapping function so that it can be compared to the subjective assessment MOS value, which includes wideband speech communications covering audio bandwidth 50 – 7000 Hz.

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.3669 \times x + 3.8224}}$$

Here,

X is the WB-PESQ value.

This mapping function is used to calculate the MOS value.

10.2. Additional considerations for measurements

(1) Input/output method of electronic signals

(i) Measurements with electronic input and output

If the following conditions are met, input/output method using handset terminals (with modular connectors) shall be provisionally recognized.

- The impedance and signal level of the equipment seen by the telephone set is equivalent to the impedance that would be presented by a normal handset.
- The signal levels to/from the equipment at the jack are equivalent to those that are present when a handset is connected to the telephone.

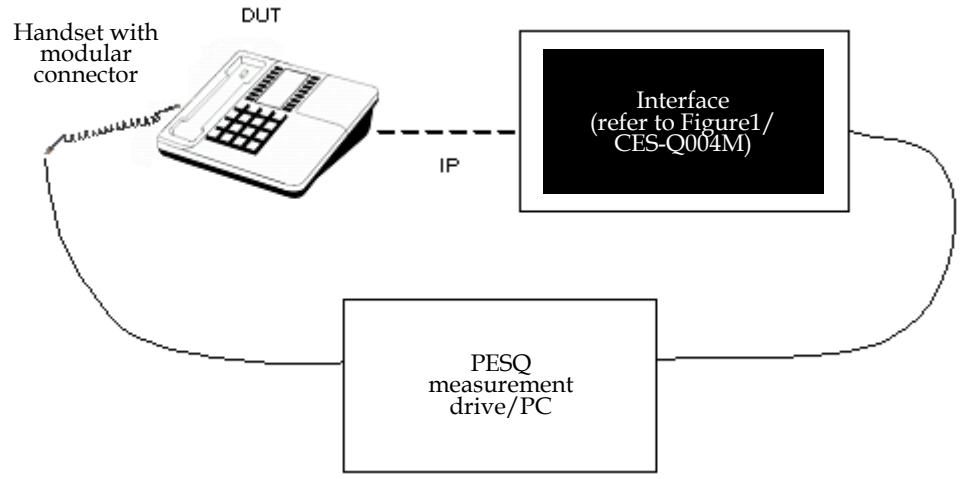


Diagram 7/Measurement method with electronic signal input/output from CES-Q005M DUT handset with modular connector

(2) Signal-to-noise ratio

The signal-to-noise ratio should be equal or greater than 30 dB when the defined level of (artificial) voice is input within the telephone bandwidth.

11. Aspects of IP disturbances

11.1. Delay variations

The probability of a certain delay occurring should follow the following exponential distribution.

The probability density distribution function of the occurrence of a delay, f(x), should be given by:

$f(x) = \lambda e^{-\lambda x}$ $\quad x \geq 0$

$f(x) = 0$ $\quad x < 0$

where λ is a parameter determining the probability of occurrence of a certain delay.

The *"delay variation"* used here in CES-Q004M-1 is the IP packet delay variation (IPDV) defined in ITU-T Rec. Y. 1541 Appendix II, and is a delay variation corresponding to an occurrence probability of $1x10^{-3}$.

For example, if IPDV is 50 ms, then the cumulative probability of the exponential distribution is

$$P(x<x_0)=\int_0^{x_0} f(x)dx=1-e^{-\lambda x}.$$

Thus, λ becomes 0.13816. Since the average of an exponential distribution is 1/λ, the average of the delay variation in this case becomes 7.238 ms. Table 2 shows the value of λ for different loading conditions.

Table 2/CES-Q003M Values and permissible error for delay variation for IP network conditions

Condition number	0	1
Width of delay variation (specified value in ms at the point where the cumulative probability is 99.9%)	-	20
Permissible error for the above (ms)	-	±4
Permissible width of delay variation (ms)	-	2.90
Permissible error for the above (ms)	-	±0.4
λ	-	0.3454

Note) Compliance should be confirmed using 100,000 or more packets.

11.2. Packet loss

It is assumed that the occurrence of packet loss is random. The bursty occurrence of packet loss is under study.

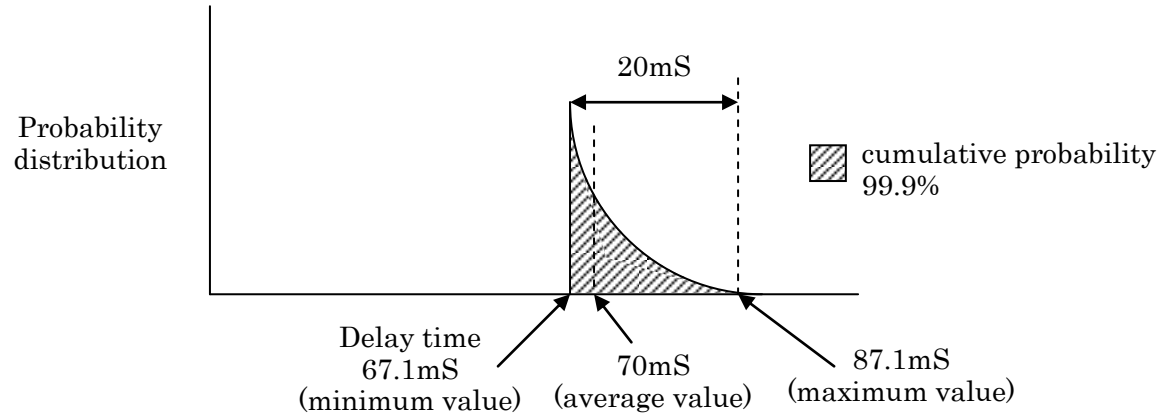
Table 3/CES-Q003M-1　Nominal values and their tolerance limits for packet loss probability for different IP network loadings

Condition number	0	1
Loss rate (%)	-	0.1
Tolerance limits	-	±0.01

Note)　Compliance should be confirmed using 100,000 or more packets.

＊）These conditions are included for terminals falling under wideband IP telephones.

The probability distribution of delay at 70 ms when exponential distribution delay variation is set at 20 ms is shown in the diagram below.



In other words, if the delay is stated as 70 ms, it is the sum of the minimum delay time (delay time with no variation) and the average value of delay variation. This is because the “delay time” measured here is the “average value,” and the “variation” is the difference from this value.