



INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.830

(02/96)

TELEPHONE TRANSMISSION QUALITY METHODS FOR OBJECTIVE AND SUBJECTIVE ASSESSMENT OF QUALITY

SUBJECTIVE PERFORMANCE ASSESSMENT OF TELEPHONE-BAND AND WIDEBAND DIGITAL CODECS

ITU-T Recommendation P.830

(Previously "CCITT Recommendation")

FOREWORD

The ITU-T (Telecommunication Standardization Sector) is a permanent organ of the International Telecommunication Union (ITU). The 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 Conference (WTSC), which meets every four years, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 1 (Helsinki, March 1-12, 1993).

ITU-T Recommendation P.830 was revised by ITU-T Study Group 12 (1993-1996) and was approved under the WTSC Resolution No. 1 procedure on the 6th of February 1996.

NOTE

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

© ITU 1996

All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the ITU.

CONTENTS

		<i>Page</i>
1	Scope	1
2	References	1
3	Definitions	2
4	Abbreviations	2
5	Conventions	3
6	Summary of subjective testing procedures	3
7	Source recordings	3
	7.1 Recording system	3
	7.2 Speech samples	4
8	Selection of experiment parameters	5
	8.1 Codec conditions	5
	8.2 Reference conditions	11
9	Experiment design	14
10	Listening test procedure	14
	10.1 Receiving system	14
	10.2 Opinion scales	14
	10.3 Electrical noise	15
11	Analysis of results	15
	Annex A – Comparison of different SNR definitions	17
	Annex B – Sets of parameters to determine codec performance	17
	B.1 Narrow-band systems (300-3400 Hz)	18
	B.2 Wideband systems (100-7000 Hz)	18
	Annex C – PCM codec corresponding to one qdu	19
	C.1 Common analogue interface	19
	C.2 Digital interface	19
	C.3 Analogue interface	19
	Annex D – Modified IRS send and receive characteristics	20
	Bibliography	21

SUMMARY

This revised Recommendation describes methods and procedures for conducting subjective performance evaluations of digital speech codecs.

Revisions encompassed by this version of the Recommendation are to include new information that reflects current practices in subjective evaluation of digital codecs, including an expanded section on creating source recordings and addition of two annexes. One annex describes an implementation of a PCM codec (A- μ -law) that generates one quantization distortion unit (qdu) of distortion to input signals. The other new annex describes the Modified IRS transmit and receive characteristics. These characteristics are recommended as the transmit and receive responses to be used in situations where the codec being tested is intended for use in fully digital circuits.

INTRODUCTION

Digital telephony in the Public Switched Telephone Network (PSTN) is now commonplace. Since the 1960s, digital transmission in the PSTN has been synonymous with 64 kbit/s A-law or μ -law PCM (see Recommendation G.711). The combination of better knowledge in signal processing techniques and advances in technology, most notably LSI/VLSI techniques, has lead to an increased interest in more efficient coding methods than PCM. Examples of this development include the introduction of 32 kbit/s ADPCM (Recommendation G.726), 16 kbit/s LDCELP (Recommendation G.728), and a 64 kbit/s, 7 kHz wideband codec (Recommendation G.722). Considerable interest is now focused on coding of speech at bit rates below 16 kbit/s.

Subjective testing is the most widely used method of assessing the performance of digital codecs. When the transmission path is digital and/or non-linear, simple objective measurements, such as those specified in Recommendation G.712, are insufficient to ensure adequate transmission performance. The aim of a subjective testing methodology is to measure the degradation contributed by the non-linear part of the transmission path, and hence to ensure that the performance of the complete system is satisfactory. To be suitable for this purpose, the measurements must be:

- a) reliable; and
- b) carried out in a way that takes account of major interactions between the non-linear part and the other parts of the transmission system.

This implies both the ability to assign a unique numerical contribution to each digital process and the ability to use this assigned contribution in conjunction with other impairments to estimate telephone connection performance.

KEYWORDS

Subjective testing, subjective evaluation.

SUBJECTIVE PERFORMANCE ASSESSMENT OF TELEPHONE-BAND AND WIDEBAND DIGITAL CODECS

(Helsinki, 1993; revised in 1996)

1 Scope

This Recommendation is based on Annex B/P.80 and on experience gained from several international experiments [5], [6], [9], [10], [12], [13], [15] and [17]. Further, this Recommendation defines a specific testing method for evaluating digital processes. This method allows the subjective effects of distortion introduced by the speech coding process to be taken into account when assessing the performance of the evolving international network. This Recommendation contains advice on how to assess the performance of digital codecs. This Recommendation must be read in conjunction with Recommendation P.80. Additional information is available in sources listed in the Bibliography.

A variety of methods are possible to characterize the performance of digital processes. At the present time these comprise listening-only tests involving:

- 1) opinion (category) ratings;
- 2) pair or multiple comparisons; and
- 3) articulation tests.

For most applications the ITU-T recommends the use of the Absolute Category Rating (ACR) method using the Listening Quality scale. However, there are times when other scales and rating methods are more suitable and appropriate (e.g. the Listening Effort scale has been found useful when interest centers on good correlation with measures of conversational performance), and these are used as well in this Recommendation. Only where there is a deviation from the use of the ACR method using the Listening Quality scale will it be stated.

NOTE – The Degradation Category Rating (DCR) method [8] is described in detail in Annex D/P.80. This method is purported to be suited when the absolute quality of the source material is low (e.g. speech corrupted by background noise) or when digital impairments are small. It may therefore be particularly useful to serve for system optimization. It should also be noted that a threshold method for direct comparison, described in detail in Annex E/P.80, is again applicable for system optimization.

Furthermore, there may be a need for supplementing listening-only tests with conversation tests planned and conducted in accordance with Recommendation P.80. However, until there has been further consideration of the circumstances in which conversational tests are necessary, and further investigation is made regarding the choice of conditions that should be covered in such experiments, no detailed plans can be included in this Recommendation.

2 References

The following Recommendations contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated are valid. All Recommendations and other references are subject to revision; all users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations listed below. A list of the currently valid ITU-T Recommendations is regularly published.

- IEC Publication 581.7 (1986), *High fidelity audio equipment and systems; Minimum performance requirements – Part 7: Loudspeakers*.
- ITU-T Recommendation G.113 (1996), *Transmission impairments*.
- ITU-T Recommendation G.191 (1993), *Software tools for speech and audio coding standardization*.

¹⁾ Previously P.83.

- ITU-T Recommendation G.192 (1996), *A common digital parallel interface for speech standardization activities.*
- CCITT Recommendation G.711 (1988), *Pulse Code Modulation (PCM) of voice frequencies.*
- CCITT Recommendation G.712 (1992), *Transmission performance characteristics of pulse code modulation.*
- CCITT Recommendation G.722 (1988), *7 kHz audio-coding within 64 kbit/s.*
- CCITT Recommendation G.726 (1990), *40, 32, 24 and 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM).*
- CCITT Recommendation G.728 (1992), *Coding of speech at 16 kbit/s using low-delay code excited linear prediction.*
- ITU-T Recommendation O.41 (1994), *Psophometer for use on telephone-type circuits.*
- CCITT Recommendation O.131 (1988), *Quantizing distortion measuring equipment using a pseudo-random noise test signal.*
- CCITT Recommendation O.132 (1988), *Quantizing distortion measuring equipment using a sinusoidal test signal.*
- ITU-T Recommendation O.133 (1993), *Equipment for measuring the performance of PCM encoders and decoders.*
- ITU-T Recommendation P.10 (1993), *Vocabulary of terms on telephone transmission quality and telephone sets.*
- CCITT Recommendation P.48 (1988), *Specification for an intermediate reference system.*
- ITU-T Recommendation P.50 (1993), *Artificial voices.*
- ITU-T Recommendation P.56 (1993), *Objective measurement of active speech level.*
- ITU-T Recommendation P.80 (1993), *Methods for subjective determination of transmission quality.*
- ITU-T Recommendation P.810 (1996), *Modulated Noise Reference Unit (MNRU).*
- ITU-T Recommendation P.84 (1993), *Subjective listening test method for evaluating digital circuit multiplication and packetized voice systems.*
- CCITT Recommendation Q.35 (1988), *Technical characteristics of tones for the telephone service.*

3 Definitions

For the purposes of this Recommendation, the following definitions apply:

- 3.1 **dBov**: dB relative to the overload point of a digital system.
- 3.2 **Q**: The ratio, in dB, of speech power to modulated noise power in the Modulated Noise Reference Unit, as described in Recommendation P.810.
- 3.3 **Q_N**: Q for a narrow-band Modulated Noise Reference Unit.
- 3.4 **Q_W**: Q for a wideband Modulated Noise Reference Unit.

4 Abbreviations

For the purposes of this Recommendation, the following abbreviations are used:

ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
BER	Bit Error Rate
DCR	Degradation Category Rating
eif	equipment impairment factor

IRS	Intermediate Reference System (Recommendation P.48)
ITU-T STL	ITU-T Software Tools Library
LDCELP	Low-Delay Code Excited Linear Prediction
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
PCM	Pulse Code Modulation
PSTN	Public Switched Telephone Network
qdu	quantization distortion unit
RMS	Root Mean Square
SNR	Signal-to-Noise Ratio

5 Conventions

Subjective evaluation of digital speech codecs may, in principle, be conducted using listening-only or conversational methods of subjective testing. As a practical matter, listening-only tests are the only feasible method of subjective testing during the development of speech codecs when a real-time implementation of the codec is not available. This Recommendation discusses listening-only methods of subjective evaluation of speech codecs.

6 Summary of subjective testing procedures

When performing a subjective evaluation of a speech codec, consideration must be given to the intended application of the codec under test. The system used for recording speech materials, the selection of experimental test conditions, and the system used for delivery of processed speech materials to the listeners must be chosen so that they capture the important end-to-end characteristics of the telephone connection. In particular, the characteristics of the recording and receiving systems should reflect the characteristics of the transmitter and receiver that will be used in the intended application of the codec.

Performing subjective evaluations of digital codecs proceeds via a number of steps:

- 1) Preparation of source speech materials, including recording of talkers;
- 2) Selection of experimental parameters to exercise the features of the codec that are of interest;
- 3) Design of the experiment;
- 4) Selection of a test procedure and conduct of the experiment;
- 5) Analysis of results.

Each of these steps is described below.

7 Source recordings

7.1 Recording system

As noted above, it is important to use a representative system for the transmitting end of the system under test. When testing narrow-band (300-3400 Hz) digital codecs for network applications, the ITU-T recommends the use of a modified Intermediate Reference System (IRS) transmit characteristic. Such a device is constructed by removing the SRAEN filter from the IRS transmitter. The transmit characteristic of the modified IRS is found in Annex D. A software implementation of the modified IRS transmit characteristic is available in the ITU-T Software Tools Library (ITU-T STL).

NOTE – The ITU-T Software Tools Library contains a number of tools that are useful for the development of speech and audio coding standards. These tools are also useful when preparing materials for use in subjective evaluations of speech codecs. The tools that are available in the ITU-T STL are described in Recommendation G.191.

In some circumstances (e.g. when it is expected that low frequencies will be sharply limited), it will be necessary or desirable to assess the performance of the codec using the sending end of the Intermediate Reference System (IRS) conforming to Recommendation P.48 and calibrated to Recommendation P.64. However, it has been found that the performance of low-bit rate speech codecs may depend significantly on the frequency characteristic applied to the input speech signal. In particular, the IRS is representative of analogue telephone connections (1970s). Hence, the modified IRS transmit characteristic should be used under most circumstances.

If a wideband system (100-7000 Hz) is to be used for audio-conferencing, then the sending end should conform to IEC Publication 581.7.

7.2 Speech samples

Speech material should consist of simple, short, meaningful sentences. These sentences should be chosen so as to be easy to understand (from the current non-technical literature or newspapers, for example). Further, the sentences should be made into sets of two or three in such a way that there is no obvious connection of meaning between the sentences in a set. Very short and very long sentences should be avoided, the aim being that each sentence when spoken should have a duration of 2-3 seconds.

NOTE – When considering bit errors, this allows, for example, for Bit Error Rates (BERs) of 0, 1:10000, 1:1000 and 1:100 to be assessed (see 8.1.4). BERs below 1:10000 will require more sentences/sample.

7.2.1 Production of source material

The talkers should be seated (one at a time) in a room with reverberation time less than 500 milliseconds, and room noise level below 30 dBA with no dominant peaks in the spectrum. The recordings should be made on a high quality recording system.

Talkers should pronounce the sentences fluently, but not dramatically, and maintain a constant level that they find comfortable, avoiding noise from sources such as the rustling of paper. The sentences may be presented to the talker one by one on a visual display controlled by a computer, or on a list at a distance suitable for easy reading.

The active speech level should be observed during recording. Care should be taken during the recording process that the active speech level for the recording is between 20 and 30 dB below the peak overload point of the recording system for each sentence measured separately. Any sentence for which this does not hold should be re-recorded.

NOTE – Care must be taken while recording, that no spurious noises (e.g. mains hum or switching spikes) are added.

7.2.2 Equalization of speech level

The measurement of signal levels in the digital part of the international network is normally expressed by telecommunications engineers as y dBm0, i.e. the level relative to 1 mW in 600 Ω. However, speech materials recorded digitally by computer are typically stored in integer format. It is more convenient to specify the levels of such material relative to the maximum signal power that can be stored on a computer, e.g. z dBov (where the characters 'ov' arbitrarily mean digital overload signal level). It is necessary to understand the relationship of these terms so that the levels in dBm0 can be expressed in dBov. Such a relationship has the form:

$$z \text{ dBov} = y \text{ dBm0} + C \quad (7-1)$$

For an A-law encoded system, a sinewave that excites the maximum level has a power given by $T_{\text{max}} = 3.14 \text{ dBm0}$ (Recommendation G.711). Similarly, for a μ-law encoded system, $T_{\text{max}} = 3.17 \text{ dBm0}$.

The reference power for dBov is the maximum integer value for the system in use. A sinewave of maximum amplitude (e.g. 32 767 for a 16-bit system) will have power -3.01 dBov . Thus, $C = -6.15 \text{ dB}$ for A-law systems and $C = -6.18 \text{ dB}$ for μ-law systems.

A typical nominal value for mean active speech level (measured according to Recommendation P.56) is -20 dBm0, corresponding to approximately -26 dBov. Thus, the active speech level of each sentence should be measured, and the recorded level adjusted to be -27 ± 1 dBov for the digital system, i.e. in digital format. For languages or voices where the peak-to-average level exceeds 23 dB, the average level of the speech material should be reduced appropriately. The software implementation of Recommendation P.56, as found in the ITU-T STL, should be used to determine the average speech level while active.

Finally, groups of sentences are combined (with a suitable silence period between sentences) to produce the final source material for use in the subjective test.

8 Selection of experiment parameters

Experimental conditions are conveniently divided into those that involve the codec under test and those that serve as references. Codec conditions provide essential information about the performance of the codec, while reference conditions allow sensible comparisons to be made when the same codec is tested in different laboratories or in the same laboratory at different times. Figures 1 to 3 and 8 to 9 are examples to demonstrate typical effects on MOS of experiment parameters.

8.1 Codec conditions

Test conditions described in this subclause will serve to indicate how the performance of the codec is affected by changes in input and listening levels, different talkers, presence of background noise, etc.

8.1.1 Speech input levels

It is recognized that digital speech codecs are sensitive to the level of the input speech and that the performance of the codec may depend on this level. When assessing the performance of codecs it is usual to set the input levels to be relative to the overload point of the codec. It is recommended that input levels of 14, 26 and 38 dB below the overload point of the codec (approximately equivalent to -8 , -20 , -32 dBm0) should be used for assessing the impact of input level variations on the performance of the codec. This is approximately equal to the mean value measured at the International Switching Point $\pm$ two standard deviations. Other applications may require different input levels. Figure 1 illustrates the effect on Mean Opinion Score (MOS) of change in speech input level.

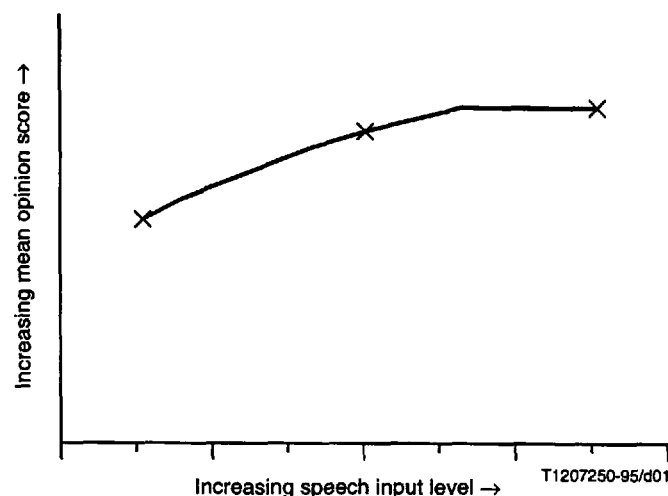


FIGURE 1/P.830
Mean opinion scores for speech input levels