



INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.83

(03/93)

TELEPHONE TRANSMISSION QUALITY SUBJECTIVE OPINION TESTS

SUBJECTIVE PERFORMANCE ASSESSMENT OF TELEPHONE-BAND AND WIDEBAND DIGITAL CODECS

ITU-T Recommendation P.83

(Previously "CCITT Recommendation")

FOREWORD

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

ITU-T Recommendation P.83 was prepared by the ITU-T Study Group XII (1988-1993) and was approved by the WTSC (Helsinki, March 1-12, 1993).

NOTES

1 As a consequence of a reform process within the International Telecommunication Union (ITU), the CCITT ceased to exist as of 28 February 1993. In its place, the ITU Telecommunication Standardization Sector (ITU-T) was created as of 1 March 1993. Similarly, in this reform process, the CCIR and the IFRB have been replaced by the Radiocommunication Sector.

In order not to delay publication of this Recommendation, no change has been made in the text to references containing the acronyms "CCITT, CCIR or IFRB" or their associated entities such as Plenary Assembly, Secretariat, etc. Future editions of this Recommendation will contain the proper terminology related to the new ITU structure.

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

© ITU 1994

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 Introduction	1
2 Source recordings	2
2.1 Sending system	2
2.2 Speech samples	2
3 Selection of experiment parameters	2
3.1 Codec conditions	2
3.2 Reference conditions	7
4 Experiment design	9
5 Listening test procedure	9
5.1 Receiving system	10
5.2 Opinion scales	10
5.3 Electrical noise	10
6 Analysis of results	10
Annex A – Comparison of different SNR definitions	12
Annex B – Sets of parameters to determine codec performance	13
B.1 Narrow-band systems (300-3400 Hz)	13
B.2 Wideband systems (100-7000 Hz)	13
References	14
Bibliography	14

Recommendation P.83**SUBJECTIVE PERFORMANCE ASSESSMENT OF TELEPHONE-BAND
AND WIDEBAND DIGITAL CODECS***(Helsinki, 1993)***1 Introduction**

Digital telephony is now being introduced in the Public Switched Telephone Network (PSTN) at an increasing speed. Since the 1960s, digital transmission in the PSTN has been synonymous to 64 kbit/s, A-law PCM (see Recommendation G.711). Now 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. One example of this development is the introduction of 32 kbit/s ADPCM (see Recommendation G.726). More recently a 64 kbit/s, 7 kHz wideband codec has been standardized by the CCITT (see Recommendation G.722). Considerable interest is now focused on Low Rate Encoding (LRE) i.e. at bit rates of 16 kbit/s and/or below.

When the transmission path is digital and/or non-linear, simple objective measurements such as 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, so that the performance of the complete system is satisfactory. To be suitable for this purpose, the measurements have to 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 uniquely assign a numerical contribution to each digital process and the ability to use this assigned contribution in conjunction with other impairments to estimate telephone connection performance.

This Recommendation, based on Annex B/P.80 and experience gained from several international experiments [2, 3, 4, 5, 6], defines a specific testing method for evaluating digital processes in a manner such that the quantization distortion effects of these processes on transmission performance can be taken into account in the evolving international network.

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 CCITT 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 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 [1] is described in detail in Annex D/P.80. This method is purported to be suited 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.

This Recommendation contains advice on how to access the performance of digital codecs. This Recommendation must be read in conjunction with Recommendation P.80.

Furthermore, there is the need for supplementing listening-only tests with conversation tests but as yet no method can be agreed for inclusion in this Recommendation.

2 Source recordings

This clause is the same as B.1/P.80 with the following exceptions.

2.1 Sending system

For narrow-band systems (300-3400 Hz) it is recommended that if comparison of results is required between laboratories and administrations that the sending end of the Intermediate Reference System (IRS) conforming to Recommendation P.48 and calibrated to Recommendation P.64 be used. 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.

NOTE –The IRS is representative of analogue telephone sets (1970s) and Study Group 12 is currently studying a “Reference Telephone” that is representative of modern telephone sets (see Figure 1/P.31).

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.

2.2 Speech samples

It is recommended that a minimum of two sentences per sample is used. This allows, for example, for bit-error-rates (BERs) of 0, 1:10000, 1:1000 and 1:100 to be assessed (see 3.1.4). BERs between 0 and 1:10000 will require more sentences/sample.

3 Selection of experiment parameters

3.1 Codec conditions

3.1.1 Speech input levels

When assessing 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 (equivalent to –8, –20, –32 dBm0) should be used. 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 demonstrates the effect of change in speech input level.

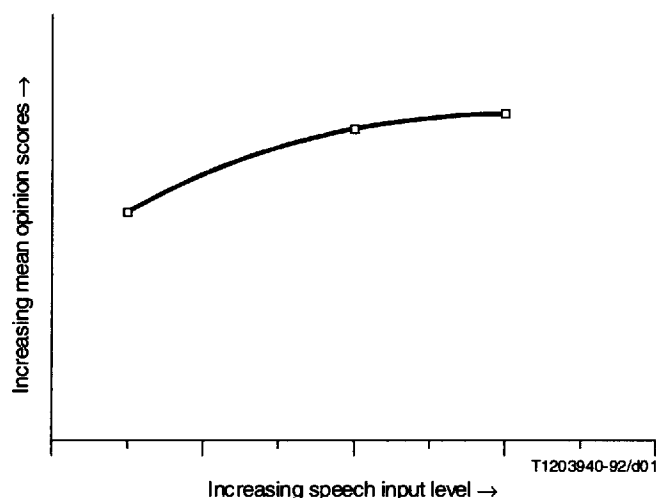


FIGURE 1/P.83
Mean opinion scores for speech input levels

3.1.2 Listening levels

It is recommended that at least three listening levels should be used. These levels should lie 10 dB either side of the preferred listening level (see B.2.1/P.80 for information on preferred listening level). Figure 2 demonstrates the effect of change in listening level.

3.1.3 Talkers

3.1.3.1 Different talkers

It is recommended that a minimum of two male and two female talkers be used. Figure 2 demonstrates the effect of different talkers. However, if talker dependency is to be tested, it is recommended to use the following talkers:

- 8 male;
- 8 female;
- 8 children.

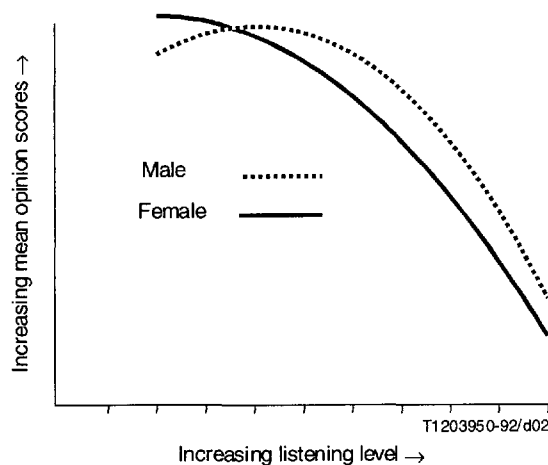


FIGURE 2/P.83
Mean opinion scores for listening levels

3.1.3.2 Multiple talkers

It is important to know the robustness of the codec to multiple voice input signals and to find out whether any adverse effects occur such as “break-up” of transmission, spurious signals, etc.

If only handset use is to be considered, then the mixing of two different talkers with input speech levels 20 dB apart is normally sufficient.

However, if it is to be used in a conference mode, i.e. hands-free, where more than one talker can speak at the same time, then one must ensure that the encoding algorithm can deal with multiple talkers where the difference in speech levels could be zero.

Either the Degradation Category Rating (DCR) method using the 5-point scale of 5.2.3 (see also Annex D/P.80) or the Quantal-Response detectability method using the 3-point scale of 5.2.4 (see also Annex C/P.80) are recommended for assessing the effects of multiple talkers.

3.1.4 Errors

This will depend on the application, i.e. whether it is used in the normal ordinary PSTN (i.e. line systems) or in say, a radio environment such as mobile radio.