



INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.84

(03/93)

TELEPHONE TRANSMISSION QUALITY SUBJECTIVE OPINION TESTS

SUBJECTIVE LISTENING TEST METHOD FOR EVALUATING DIGITAL CIRCUIT MULTIPLICATION AND PACKETIZED VOICE SYSTEMS

ITU-T Recommendation P.84

(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.84 was revised 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 Similarly, in this reform process, the CCIR and the IFRB have been replaced by the Radiocommunication Sector.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

	Page
1 Introduction	1
1.1 Purpose	1
1.2 Test philosophy.....	1
2 Source recordings.....	2
2.1 Apparatus and environment.....	2
2.2 Speech material.....	2
2.3 Procedure	3
2.4 Calibration signals and speech levels	4
3 Simulating system load	4
3.1 Requirements for a generic voice load simulator	4
3.2 Determining load capacity of tested systems.....	5
3.3 Controlling load applied to tested systems	5
4 Processing of the speech	6
5 Test design	6
5.1 Test No. 1 – Effect of applied load.....	8
5.2 Test No. 2 – Effect of digital errors in the DCME control channel.....	9
6 Listening test procedure	9
7 Analysis of results	10
Annex A – Description of Digital Circuit Multiplication Equipment	10
A.1 Definitions	10
A.2 Digital Speech Interpolation (DSI).....	13
A.3 Speech Detection	13
A.4 DCME load.....	13
A.5 Overload strategies	14
A.6 Silence reconstruction methods	16
A.7 Circuit versus packet mode.....	16
A.8 Packet reconstruction.....	17
Annex B – Speech material used to construct speech sequences	17
Annex C – Instructions on the use of a limited number of sentences.....	19

Recommendation P.84**SUBJECTIVE LISTENING TEST METHOD FOR EVALUATING
DIGITAL CIRCUIT MULTIPLICATION AND PACKETIZED VOICE SYSTEMS¹⁾***(Melbourne, 1988, amended at Helsinki, 1993)***1 Introduction****1.1 Purpose**

The purpose of this Recommendation is to describe a subjective listening test method which can be used to evaluate the speech quality of Digital Circuit Multiplication Equipment (DCME) or packetized voice systems. This Recommendation is intended for use with DCME systems, such as those described in Recommendation G.763, which use digital speech interpolation (DSI) techniques. For subjective test evaluations of DCME systems not using digital speech interpolation techniques, described in Recommendation G.726 [40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)], use Recommendation P.83.

Many of the degradations found in DCME or packetized voice systems have not been tested before and their effects on other systems in the network are unknown. Therefore the only definitive method is the conversation test where the effects of non-linearity, delay, echo, etc. and their interactions can be verified.

For DCME systems, degradations can include not only the effects of variable bit-rate coding, DSI gain (channel allocation), clipping, freezeout and noise contrast, but also those due to non-linearities in the speech detection system, such that the system may function differently for different speech input levels or activity factors. For packetized voice systems the subjective effect, for example, of "lost packets" is unknown.

Listening tests play an important preliminary role in the assessment, and can supply useful information service to narrow the range of conditions needing a complete conversation test. This listening test method will not provide results for generating network application rules based on factors analogous to the quantizing distortion unit (qdu). Future improvements of the test will allow such results to be obtained.

Evaluation of DCME in tandem with other DCME has not been considered at this stage, nor have the effects of systems using encoding at different rates. This Recommendation will subsequently be updated when information of these specific points becomes available.

This Recommendation confines itself solely to listening tests; a separate recommendation, on conversation tests, will be formulated when sufficient information on evaluation techniques is available. Alternatively, this Recommendation may be revised to include conversation test methods.

1.2 Test philosophy

In order for a test to satisfactorily evaluate DCME performance the test methodology should meet certain conditions. These are as follows:

- i) the method should use principles, procedures, and instrumentation that are acceptable to CCITT;
- ii) the method should be adaptable to different languages and should yield results that are comparable to those from other tests performed using this Recommendation;

¹⁾ The specifications in this Recommendation are subject to future enhancements and therefore should be regarded as provisional.

- iii) the method should permit DCME performance to be compared subjectively (or objectively) to reference conditions. Examples of suitable reference conditions are hypothetical reference connections (HRCs), white noise and speech correlated noise. The HRCs should model the facilities the DCME is designed to replace, when these facilities are known. The results of the comparisons should permit making “equivalence statements” about the DCME, e.g. a DCME system is subjectively equivalent to x asynchronously tandemed 64 kbit/s PCM systems. Ideally, the method should yield results from which a network application rule can be derived;
- iv) the DCME should be tested with a realistic load simulator and circuit-under-test signal conditions applied. Most of the transitory impairments arise when the DCME is operating in the range of applied load which forces the use of DSI. Therefore, to subjectively measure the effects of these impairments it is necessary to vary the applied load on the DCME up to and marginally beyond the maximum design load. The clipping produced by the speech detector is affected by the type of signal being transmitted on the circuit under test. Therefore, only a realistic speech signal which also contains appropriate additive noise should be used on the circuit under test;
- v) the methodology should, ideally, yield results which can be used to produce new opinion models or modify existing models.

2 Source recordings

2.1 Apparatus and environment

See B.1.1/P.80 through B.1.3/P.80.

2.2 Speech material

The format of the speech material must be suitable for the opinion scale being used in the test. This will normally be the listening-quality scale, but optionally the listening-effort scale will be used (see clause 5).

When the listening-quality scale is used, the following requirements apply:

- i) The speech material should consist of short passages (called segments), chosen at random (from current non-technical literature or newspapers for example), easy to understand, and more or less self-contained in meaning.
- ii) Each segment when spoken naturally should have a duration not less than 9 seconds and not more than 11 seconds.
- iii) Each segment should consist of at least three “sentences” in a broad sense, that is, parts that can naturally be separated by pauses in speaking, but connected in meaning to what precedes and follows within the segment.
- iv) Within each segment there should be at least one pause which naturally, in view of the meaning and structure of the text, would last for 1 to 2 seconds. The other pauses must be of natural length, since unnaturally long or short pauses may well be interpreted as reducing the quality of the speech. The simplest way of ensuring this is to construct the script of each text with either a special mark or the beginning of a new paragraph at the point where the 1 to 2 second pause is desired. The talkers recording the segments can then be instructed to read consecutively, making sure that they pause for a second or two at the marked point, and pausing naturally at other points.

When the listening-effort scale is used, the following requirements apply:

- v) The speech material should consist of single meaningful sentences, easy to understand, chosen at random (from current non-technical literature or newspapers for example) and assembled into groups (called segments). The number of sentences in every segment should be the same (three is the recommended number).
- vi) There should be no obvious connection of meaning between one sentence and another in the same segment. This precaution is to reduce the context information within the segment to a minimum, so as not to inflate the opinion scores artificially.

- vii) The pauses between the sentences must be at least one second in length. This ensures that the listeners perceive the sentences as isolated from each other in meaning, and puts the DCME system to the test in respect of the unbridged gaps in speech. Moreover, each segment should have a duration not less than 9 seconds and not more than 11 seconds including the pauses.

This structure can be attained by either:

- giving timing cues to talkers at the recording stage [see Section 2.5.8.1 d) of the *Handbook on Telephonometry*]; or
- editing the recordings afterwards.

Either of two approaches may be used for the listening-quality scale:

- i) to have as many different segments as there are conditions (an example of suitable material from which segments may be constructed is contained in Annex B); or
- ii) to have a more limited number, e.g. 10 segments per talker, where combinations of two segments can be used (this is shown in detail in Annex C). In this case, additional precautions would have to be taken in the analysis of variance of results from the tests.

The first approach is essential when the listening-effort scale is used, because listening-effort scores are affected when the listener has heard the same sentences before. Enough segments must be available to cater for all the test conditions, plus a sufficient number for use in a practice session.

2.3 Procedure

A silent period containing only circuit noise of approximately one second should proceed each segment and the segment should end with a similar silent period containing only the circuit noise.

To facilitate the processing of the recorded speech through the DCME, i.e. to allow for the starting and stopping of the recorders between segments and to allow time for adjusting the DCME for the next test condition, segments should be separated by a 5 second gap on the tape. Therefore, the recorded source segments will have the pattern on the tape shown in Figure 1 (this is an example for the listening quality scale; for the listening effort scale all pauses must be at least 1 second). Note that if the speech is stored digitally (e.g. on a disk based system) the segments will not require these 5 second gaps.

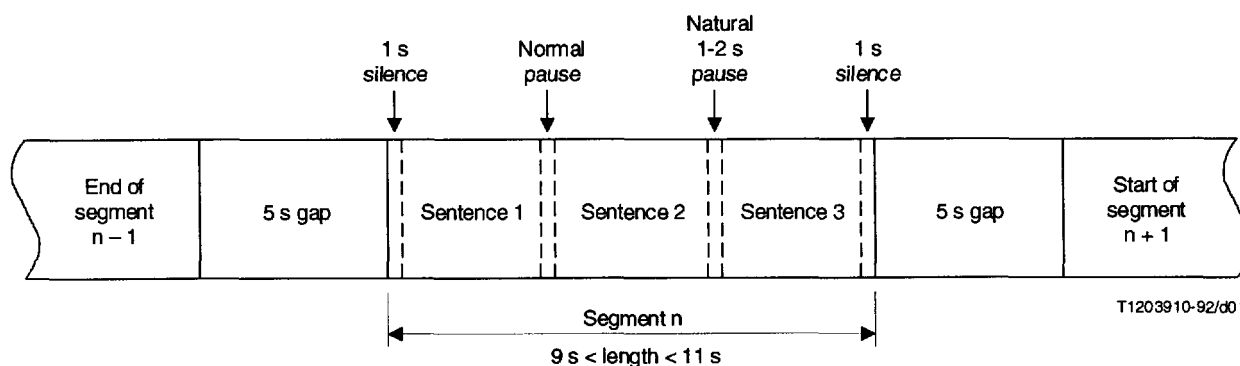


FIGURE 1/P.84

**Example of a speech segment in the format required
for the listening quality scale, containing three sentences**

The recording procedure detailed in Section 2.5.8.1 (Listening Tests) of the *Handbook on Telephonometry* should be followed. Only the part dealing with recording through an IRS is required for this Recommendation.

NOTE – When this recording is made through a physical IRS, the sidetone path of the IRS should be set to 12 dB STMR. This helps to stabilize the speech level of the talker.

Segments should be played back to listeners complete with the silent period. After the segment has ended, a sufficient length of complete silence should be provided to permit the listener to vote.

Talkers should pronounce the segment of sentences fluently but not dramatically and have no speech deficiencies such as “stutter”. (See also B.1.6/P.80.)

2.4 Calibration signals and speech levels

See B.1.7/P.80 and B.1.8/P.80.

3 Simulating system load

3.1 Requirements for a generic voice load simulator

Digital Circuit Multiplication Equipment (DCME), by definition, is used to gain an advantage in the number of circuits multiplexed onto a digital transmission facility. With this advantage, however, comes potential degradation of transmission quality when carried loads exceed that for which the DCME was engineered. Thus, a rigorous performance evaluation of DCME includes studying the behaviour of the DCME under conditions of no load, engineered load, and overload. Because the transmission performance of DCME under load depends critically upon the load characteristics, it is necessary to know and control simulated loads in order to properly assess DCME performance. This subclause describes the generic requirements for a voice load simulator for the purpose of facilitating DCME performance evaluations under conditions that are meaningful. Use of voice load simulators with the generic requirements described here will also enable the comparison of results from different studies of various DCME.

Either the load simulator or the DCME system itself must be programmed so as to record, for each individual segment played through it, the proportion of time during that segment for which the “live” channel was ON, or equivalent information. This quantity x_i will be called the concomitant variable for segment i [see Section 2.5.2 f) of the *Handbook on Telephonometry*].

NOTES

1 The load simulator specified here is to be used for the performance evaluation of DCME using Digital Speech Interpolation (DSI). This excludes Type A (see A.1), for which load is not an issue by virtue of the fixed time-slot assignment of the channels.

2 The load simulator specified here is an “external” simulator that produces simulated speech signals so as to exercise many circuits being multiplexed onto a digital transmission facility. Prototype DCME frequently use “internal” load simulation of “trunk needs service” requests that simulate the output of multiple speech detector circuits and thus compete for transmission capacity, even though no simulated signals are actually transmitted; only the “live” channel under test is actually transmitting. This internal type of simulator can be very useful in the lab, but is not treated here because certain assumptions would have to be made regarding the performance characteristics of the associated speech detector simulation.

3.1.1 Parameters

A generic Voice Load Simulator (VLS) for DCME performance evaluation has the following attributes (the parametric specification of which are detailed later in this subclause):

- talk-spurt characteristics;
- silence (gap) characteristics;
- background noise-fill for silent periods;
- spectral properties of the simulated speech;
- amplitude characteristics;
- physical interface, including idle-circuit specifications.