INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.80

(03/93)

TELEPHONE TRANSMISSION QUALITY
SUBJECTIVE OPINION TESTS

METHODS FOR SUBJECTIVE DETERMINATION OF TRANSMISSION QUALITY

ITU-T Recommendation P.80

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 Similarly, in this reform process, the CCIR and the IFRB have been replaced by the Radiocommunication Sector.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

2 Recommended methods 1
  2.1 Conversion opinion tests 1
  2.2 Listening opinion tests 1
  2.3 Interview and survey tests 2
  2.4 Other tests 3

Annex A – Conversation-opinion tests 3
  A.1 Test facilities 3
  A.2 Experiment design 7
  A.3 Conversation task 7
  A.4 Test procedure 8

Annex B – Listening tests – Absolute Category Rating (ACR) 10
  B.1 Source recordings 10
  B.2 Selection of circuit conditions 12
  B.3 Design of experiment 13
  B.4 Listening test procedure 13

Annex C – Quantal-Response Detectability Tests 16

Annex D – Degradation Category Rating (DCR) method 17
  D.1 Introduction 17
  D.2 Degradation Category Rating (DCR) procedure 17
  D.3 Statistical analysis 18

Annex E – The threshold method for comparison of transmissionsystems with a reference system 18
  E.1 Introduction 18
  E.2 Testing procedure 19
  E.3 Presentation of signals 19
  E.4 Speech sources 20
  E.5 Listening environment 20
  E.6 Listeners 20
  E.7 Reliability 20

References 20

Bibliography 21

METHODS FOR SUBJECTIVE DETERMINATION OF TRANSMISSION QUALITY

(amended at Helsinki, 1993)

1 Introduction

This Recommendation contains advice to Administrations on conducting subjective tests of transmission quality in their own laboratories. It does not however deal with types of tests described in detail in other Recommendations, namely:

a) determination of Reference and Relative Equivalents – see Recommendation P.72 (*Red Book*, Volume V, Geneva, 1985 and *Handbook on Telephonometry*, Geneva, 1987);

b) determination of Loudness Ratings – see Recommendation P.78;

c) determination of Articulation Ratings (A.E.N. values) – see Recommendation P.45 (outline in *Yellow Book*, Volume V, 1981; full detail in *Red Book,* Volume V, 1962).

Neither does it deal with the various kinds of specialized tests used in the course of developing items of telephone equipment, for the purpose of diagnosing faults and shortcomings, such as Diagnostic Rhyme Tests [1] and other tests dedicated to the study of specific aspects of speech output.

This Recommendation gives the approved methods which are considered to be suitable for determining how satisfactorily given telephone connections may be expected to perform.

The methods indicated here are intended to be generally applicable whatever the form of degradation factors present. Examples of degrading factors include: loss (often frequency dependent), circuit noise, environmental noise, sidetone, talker echo, non-linear distortion of various kinds including low-rate encoding, propagation time, harmful effects of voice-operated devices, distortions of the time scale arising from packet switching, and time-varying degradations of the communication channel, including those arising in loudspeaking sets. Combinations of two or more of such factors also have to be catered for.

References are common to all the parts and are given at the end. All references to the *Handbook on Telephonometry* are to the second edition (1993).

2 Recommended methods

2.1 Conversion opinion tests

Laboratory conversation tests are intended – as far as possible – to reproduce, in the laboratory situation, the actual service conditions experienced by telephone customers. To this end it is necessary to choose the circuit conditions and subjects suitably, and to administer the tests in an appropriate manner.

It is important that the conditions simulated in the test are correctly specified and set-up, and measured accurately before and after each experiment; that auxiliary facilities such as dialling and ringing are provided; and that faithful records of the output of each test are kept. Detailed description of the method, considerations and precautions are found in Annex A.

2.2 Listening opinion tests

Listening opinion tests are not expected to reach the same standard of realism as conversation tests, and the restrictions are therefore less severe in some respects; but the artificiality that has to be accepted brings with it a necessity for strict control of many things which in conversation tests are allowed to find their own equilibrium.

The recommended test method for listening-only tests is the "Absolute Category Rating" (ACR) method described in Annex B, which is in conformance with the Category Judgement method recommended for conversation tests (see Annex A), and adopted partly for the same reasons. Category ratings are applied to short groups of unrelated sentences, passed through a number of standard processes as well as the processes under test. This method is well established, and has been applied to both analogue telephone connections and digital codecs. In the work leading to Recommen-dations G.726 32 kbit/s ADPCM and G.722, for example, laboratories in different countries performed subjective tests by the same method on the same physical conditions and on identical transmission systems, and the results showed a high degree of consistency.

Other methods commonly used are the Quantal-Response Detectability Method, Degradation Category Rating (DCR) and the Threshold Method.

Annex C describes Quantal-Response Detectability Tests, which are suitable for evaluating threshold values of certain quantities and their associated probabilities. For example, the level above which single-frequency interference has a given probability of being objectionable or detectable, or the probability that crosstalk in a given range of levels is intelligible, can best be determined by this method.

An alternative to the Absolute Category Rating method is the Degradation Category Rating (DCR) method which is described in detail in Annex D. The DCR method compares the system under test with a high quality fixed reference and the degradation (from "Inaudible" to "Very annoying") is rated on a five-point scale. This method is suitable when the impairment (especially digital impairments) is small. It may therefore be particularly useful for evaluating e.g. algorithms for digital speech processes which are very similar and so serve as a means for system optimization, when once it has been shown by the methods of Annexes A and B that the worst-case connection incorporating the degradation in question is within acceptable limits.

The Threshold method, also suitable for system optimization, is described in Annex E. By direct comparison of the system under test with a reference system, such as the Modulated Noise Reference Unit (MNRU) as described in Recommendation P.81, it is possible to equate the value of the reference condition (Q for digital processes) which equals the performance of the system under test.

Information on other types of subjective test methods, which include scaling methods, can be found in 2.6 of the *Handbook on Telephonometry*.

One way listening tests have direct applications in the assessment of physical transmission systems which are essentially unidirectional: examples include broadcast circuits, public address systems and recorded announcement systems in which listening degradations such as loss, noise and distortion may be present.

Results of listening-only tests can be applied, but only with certain reservations, to the prediction of the assessment for conversation conducted over a both-way system, such as a connection in a public switched telephone network. The provisos are that the effects of the following additional factors are duly taken into account:

- talking degradations (e.g. sidetone and echo);
- conversation degradations (e.g. propagation time and mutilation of speech by the action of voice-operated devices).

The following subclauses deal with: preparation of speech material; processing of speech material; experiment philosophy, which includes choice of circuit conditions; listening test procedure and treatment of results.

2.3 Interview and survey tests

If the rather large amount of effort needed is available and the importance of the study warrants it, transmission quality can be determined by "service observations". Recommended ways of performing these, including the questions to be asked when interviewing customers, are given in Recommendation P.82. To maintain a high degree of precision a total of at least 100 interviews per condition is required.

A disadvantage of the service-observation method for many purposes is that little control is possible over the detailed characteristics of the telephone connections being tested. However, this method does afford a global appreciation of how the "equipment" performs in the real environment.

Further information can be found in the *Handbook on Telephonometry*, 2.5.8.3.

2.4 Other tests

Reference [2] gives information of a method that largely overcomes the disadvantages of the interview technique of 2.3 yet retains many of the advantages. This method, termed SYBIL, allows a small proportion of their ordinary calls to be passed through special arrangements which modify the normal quality of transmission according to a test programme. If a particular call has been so treated the volunteer is asked to vote by dialling one of a set of digits to indicate his opinion. In this way all results are recorded by the controlling computer and complete privacy is maintained.

Annex A

Conversation-opinion tests

(This annex forms an integral part of this Recommendation)

A.1 Test facilities

A.1.1 Physical conditions

A.1.1.1 Test cabinets

The two subjects are seated in separate sound-proof cabinets near the point from which the experiment is controlled. The volume of the room is not less than 20 m^3, with a reverberation time less than 500 ms (normally in the range of 200-300 ms), for handheld systems such as telephone handsets, or for headset systems; and not less than 30 m^3 for handsfree systems (extra care is exercised if reverberation time is an experimental variable).

The internal dimensions of the cabinet are such that standing-wave pattern effects are kept to a minimum. A typical ratio is 5:4:3.

The physical construction of the rooms should be such that sufficient sound attenuation of the outside noise environment is achieved so that the requirements of A.1.1.2.1 are met.

The cabinets are favourably decorated to recreate a natural environment.

A.1.1.2 Noise

A.1.1.2.1 Noise floor

The ambient noise level (when no environmental noise is deliberately set up) is kept as low as possible. For practical reasons such as regular changes of fresh air in the cabinet, the target is an upper limit of NC25 [3] or NR25 [4]. These values approximate to the noise level in homes (sleeping areas), hospitals and libraries.

A.1.1.2.2 Environmental noise

Environmental noise is fed in with the required spectrum (e.g. Hoth spectrum to represent typical room noise – see A.1.1.2.2.1) at the required level (e.g. 50 dBA) measured with a Precision Sound Level Meter conforming to IEC Publication 651:1979, used with the “A weighting” and the “fast” meter characteristic. If different conversations in the same experiment require different room noise levels, then care is taken to prevent the transitions from being too obvious to the subjects. Ideally, room noise should be changed only when subjects are out of the sound-proof rooms. If this is not possible, then changes of level are carried out gradually (at a rate not exceeding 4 dB per second), at a time when no experimental conversation is in progress and when the subjects’ attention is otherwise occupied – by communicating with the operator, for example.

Spectra with long-term characteristics are given in A.1.1.2.2.1 and A.1.1.2.2.2.

For some purposes it is necessary to use noise that fluctuates in level or spectrum, such as tape recordings of actual office noise or traffic noise. In such cases it should be ensured that the statistical characteristics are stable when averaged over a reasonably short period of time such as one minute.

It is recommended that the noise level and spectrum are measured at least twice; at the beginning and end of the experiment. Any significant variation in the two measurements, when compared with each other, must be assessed by the experimenter as it may cast doubt on the validity of the experiment.

It is essential to ensure that the loudspeakers and amplifiers are capable of faithfully reproducing the required room noise.

A.1.1.2.2.1 Room noise

The room noise shall have a power density spectrum corresponding to that published by Hoth [5]. Table A.1 gives the spectrum density adjusted in level to produce a reading of 50 dBA on a sound level meter conforming to IEC Publication 179 [6]. This is produced in Figure A.1. This spectrum is independent of level, i.e. for 40 dBA the level in each band shall be 10 dB less than that shown in Table A.1. Additional information on the power in each one-third octave band is also given in Table A.1.

TABLE A.1/P.80

Room noise spectrum

Frequency (Hz)	Spectrum density (dB SPL/Hz)	Bandwidth 10 log$_{10}$ Δf (dB)	Total power in each 1/3rd octave band (dB SPL)	Tolerance (dB)
100	32.4	13.5	45.9	±
125	30.9	14.7	45.4	
160	29.1	15.7	44.9	
200	27.6	16.5	44.1	
250	26.0	17.6	43.6	
315	24.4	18.7	43.1	
400	22.7	19.7	42.3	
500	21.1	20.6	41.7	
630	19.5	21.7	41.2	
800	17.8	22.7	40.4	
1000	16.2	23.5	39.7	
1250	14.6	24.7	39.3	
1600	12.9	25.7	38.7	
2000	11.3	26.5	37.8	
2500	9.6	27.6	37.2	
3150	7.8	28.7	36.5	
4000	5.4	29.7	34.8	
5000	2.6	30.6	33.2	
6300	–1.3	31.7	30.4	
8000	–6.6	32.7	26.0	

NOTES

1 The electrical input signal, e.g. white noise, shall be band-limited to the 1/3rd octave bands centred on the ISO preferred frequencies (ISO 266) between 100 Hz and 8000 Hz with the band edges conforming to the filters described in IEC 225.

2 The acoustical room noise is difficult to control at low frequencies, especially in the unspecified region below 100 Hz because of the dimensions of typical test cabinets, poor attenuation of such cabinets and the influence of extraneous noises, e.g. air-conditioning plant. It is therefore desirable to select a test cabinet that keeps these unwanted low frequency sound pressure levels to a minimum.

A.1.1.2.2.2 Internal vehicle noise

Two spectra representing internal vehicle noise [7], [8] are recommended. They are adequately represented by simplified curves [9]: one spectrum for moving vehicles and the other for stationary vehicles. Table A.2 gives the spectrum densities together with additional information on the power in each one-third octave band. The spectrum density for moving vehicles is shown in a) of Figure A.2 and for stationary vehicles in b) of Figure A.2. These spectra are independent of level.

Table A.3 gives the computed values of the unweighted sound pressure levels for various speeds calculated over the ISO one-third octave frequency bands centred on 63 Hz to 8000 Hz.

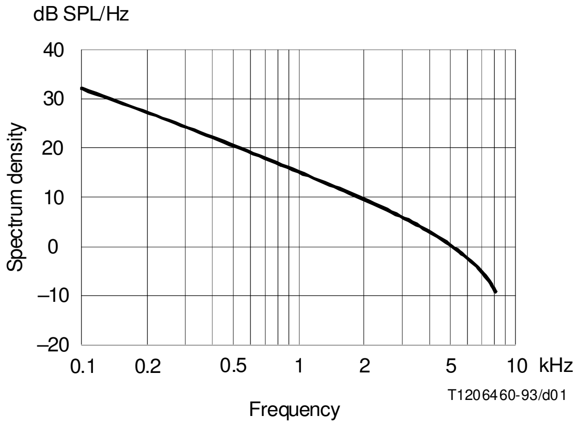


FIGURE A.1/P.80

Room noise spectrum density

TABLE A.2/P.80

Internal vehicle noise spectra

Frequency (Hz)	Spectrum density (dB SPL/Hz)		Bandwidth $10 \log_{10} \Delta f$ (dB)	Total power in each 1/3rd octave band (dB SPL)		Tolerance (dB)
	Moving	Stationary		Moving	Stationary	
63	72.3	58.3	11.7	84.0	70.0	±
80	69.3	55.0	12.7	82.0	66.7	
100	66.5	49.8	13.5	80.0	63.3	
125	63.3	45.1	14.7	78.0	60.0	
160	60.3	42.0	15.7	76.0	56.7	
200	57.5	36.8	16.5	74.0	53.3	
250	54.4	34.7	17.6	72.0	52.3	
315	51.3	32.6	18.7	70.0	51.3	
400	48.3	30.6	19.7	68.0	50.3	
500	45.4	28.7	20.6	66.0	49.3	
630	42.3	26.6	21.7	64.0	48.3	
800	39.3	24.6	22.7	62.0	47.3	
1000	36.5	22.8	23.5	60.0	46.3	
1250	33.3	20.6	24.7	58.0	45.3	
1600	30.3	18.6	25.7	56.0	44.3	
2000	27.5	16.8	26.5	54.0	43.3	
2500	24.4	14.7	27.6	52.0	42.3	
3150	21.3	12.6	28.7	50.0	41.3	
4000	18.3	10.6	29.7	48.0	40.3	
5000	15.4	8.7	30.6	46.0	39.3	
6300	12.3	6.6	31.7	44.0	38.3	
8000	9.3	4.6	32.7	42.0	37.3	

TABLE A.3/P.80

Computed sound pressure levels of spectra

Spectra		Sound pressure level, unweighted (dB SPL)
Moving	30 km/h	80
	80 km/h	85
	110 km/h	90
Stationary		75

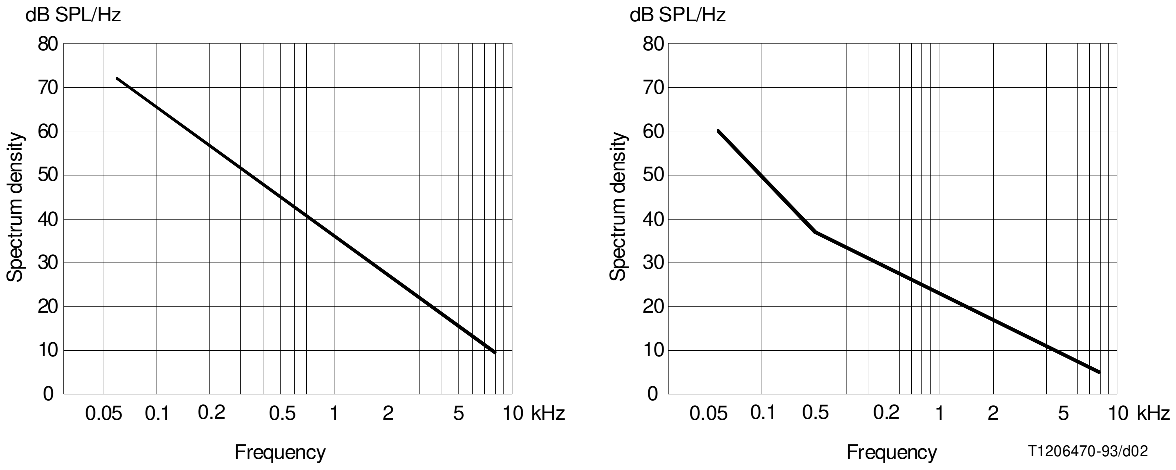
NOTES to Tables A.2 and A.3:

1 These values apply for typical vehicles. Discretion may be used to adjust the levels downwards for luxury vehicles and upwards for noisier vehicles.

2 Because of the practical difficulty of generating such high sound pressure levels at low frequencies, and because normal speech contains no apparent energy below about 63 Hz in which range of frequencies the car is also comparatively intensitive it is probably advisable to restrict the recommended noise spectrum to frequencies above 63 Hz. However, it should be borne in mind that low and medium frequency vibrations have important physiological and psychological effects which should be studied in their own right.

3 The electrical input signal, e.g. white noise, shall be band-limited to the 1/3rd octave bands centred on the ISO preferred frequencies (ISO 266) between 63 Hz and 8000 Hz with the band edges conforming to the filters described in IEC 225.

4 The acoustical room noise is difficult to control at low frequencies especially in the unspecified region below 63 Hz because of the dimensions of typical test cabinets, poor attenuation of such cabinets and the influence of extraneous noises e.g. air-conditioning plant. It is therefore desirable to select a test cabinet that keeps these unwanted low frequency sound pressure levels to a minimum.



a) Spectrum density for moving vehicles

b) Spectrum density for stationary vehicles

FIGURE A.2/P.80

Vehicle noise spectral density

A.1.1.3 Noise measurement position

It is recommended that the measurement of sound pressure level (SPL) in the test cabinets (see A.1.1.1) shall be made as follows:

- furniture in position;
- no subject or test personnel present;
- the SPL shall be measured at a vertical distance of 740 mm above the centre of the seat of the subjects chair with a meter conforming to Recommendation P.54 using "A" weighting;
- the spectrum of the environmental noise shall be measured in one-third octaves, centred at the preferred frequencies as defined in ISO R.266, and must stay within the specified tolerances e.g. ± 3 dB for Hoth noise (see A.1.1.2.2.1);
- in rooms where more than one subject is to be tested the difference in dBA, for all subject positions, shall not vary by more than ± 2 dB.

NOTE – It is suggested that the minimum distance between each loudspeaker and the measurement position should be 1.5 m.

A.1.2 Establishing the connection

When establishing the laboratory connection, the following must be taken into consideration:

- telephone sets;
- initial call set-up;
- laboratory representation of telephone connections.

It is recommended that the sensitivity/frequency characteristic of the connection is measured at least twice; at the beginning and end of the experiment. Any significant variation in the two measurements, when compared with each other, must be assessed by the experimenter as it may cast doubt on the validity of the experiment.

Detailed information on this aspect can be found in the *Handbook on Telephonometry*, 2.5.8.2.

A.1.3 Monitoring

Monitoring can take many forms but the three most commonly used are:

- *Intercommunication system* – Essential to allow the subject and experimenter to communicate with each other.
- *Visual monitoring* – This has two purposes; the first being safety and the second to observe the peculiarities of the subject, e.g. how they hold the telephone handset.
- *Tape recordings and recording system* – This facilitates other useful information to be gleaned, e.g. duration of the call, speech voltage, speech activity.

Detailed information on this aspect can be found in the *Handbook on Telephonometry*, 2.5.8.2.

A.2 Experiment design

There are a number of methods suitable for use in designing experiments, e.g. Latin Squares, Youden Squares, Balanced Incomplete Blocks and Randomization with Replication to name but a few. The experimenter must decide for himself the method to be used taking into account the number of test conditions, accuracy of results and the ability to make sound judgements from the findings.

The most suitable designs are of the $n \times n$ graeco-latin square type and a detailed description can be found in the *Handbook on Telephonometry*, 2.5.8.2.

A.3 Conversation task

Every effort is to be made to ensure that conversations are purposeful, and that subjects have full opportunity to exploit the transmission capabilities of the test circuit.

The general rule is that every conversation should have a natural beginning and a natural ending. Unless it is absolutely necessary, the conversation must never be terminated in the middle of the task (with the exception of Simplified Conversation Tests which are described in the *Handbook on Telephonometry*, 2.5.8.2).

An example of a conversational task can be found in the *Handbook on Telephonometry*, 2.5.8.2d.

A.4 Test procedure

A.4.1 Eligibility of subjects

Subjects taking part in the conversation tests are chosen at random from the normal telephone using population, with the provisos that:

a) they have not been directly involved in work connected with assessment of the performance of telephone circuits, or related work such as speech coding; and

b) they have not participated in any subjective test whatever for at least the previous six months, and not in a conversation test for at least one year.

If the available population is unduly restricted, then allowance must be made for this fact in drawing conclusions from the results.

No steps are taken to balance the numbers of male and female subjects unless the design of the experiment requires it. Subjects are arbitrarily paired in the experimental design prior to the test and remain thus paired for its duration.

A.4.2 Opinion scales

The following opinion scales are those recommended by the CCITT.

A.4.2.1 Conversation opinion scale

Various five-point category-judgement scales may be used for different purposes. The layout and wording of opinion scales, as seen by subjects in experiments, is very important, and should follow the standard arrived at through years of experience. The following opinion scale is the most frequently used for CCITT applications and equivalent wording should be used depending on language which might result in small variations to the original English text.

This is a category rating obtained from each subject at the end of each conversation.

Opinion of the connection you have just been using

Excellent

Good

Fair

Poor

Bad

The experimenter allocates the following values to the scores:

Excellent = 5 Good = 4 Fair = 3 Poor = 2 Bad = 1

and all further statistical processing is performed in terms of these numbers. The arithmetic mean of any collection of these opinion scores is called **mean conversation-opinion score**, and is represented by the symbol Y_C (or, where suffix notation is not available, the symbol Y_c).

NOTES

1 In the past, the equivalences Excellent = 4, Good = 3, Fair = 2, Poor = 1, Bad = 0 have often been used. Anyone using results from earlier experiments must be aware that the mean scores must all be increased by one to be comparable with those now obtained; otherwise there is no difference in the numerical processing that may be applied.

2 This scale is labelled in English and the equivalent should be inserted depending on language which might result in small variations to the original English text.

A.4.2.2 Difficulty scale

This is a binary response obtained from each subject at the end of each conversation.

Did you or your partner have any difficulty in talking or hearing over the connection?

Yes

No

The experimenter allocates the following values to the scores:

Yes = 1 No = 0

The quantity evaluated (percentage of "yes" responses) is called percentage Difficulty or per cent "Difficult", and is denoted by the symbol %D. The corresponding simple proportion is denoted by the symbol d; in other words, %D = 100d.

NOTE – It is often the case that the nature of the difficulty is required and then it is usual for the experimenter to ask the subject to describe in his/her own words their perception of the difficulty.

The layout and wording of the opinion scale, as seen by subjects in experiments, are very important, and should follow the standard arrived at through years of experience: see A.4.3.

A.4.2.3 Other opinion scales

These are variants of the methods of "magnitude estimation" and "cross-modality matching" [10] and their responses shall be one of the following:

a) one of a numerical series of categories labelled 1, 2, 3, 4, 5 (and denoted as such to the subject), but with descriptions attached only to the first and the last, to identify the subjective dimension;

b) a numerical mark on a scale from one to a number much greater than five – say 10 or 100; or

c) a length proportional to some property (e.g. quality), marked manually along a given straight line.

A survey of experimental methods can be found in 2.6.2 of the *Handbook on Telephonometry*.

A.4.3 Instructions to subjects

Instructions are given to subjects on arrival for their first visit. It is normal for the subject to receive a letter prior to arrival which contains non-technical information on the experiment and what will be expected of them. An example of such a letter can be found in 2.5.8.2, Table 3/2.5, of the *Handbook on Telephonometry*.

They are asked whether they have read and understood the letter. Any obscurities are clarified, and opportunity is given for asking questions. The sound-proof rooms and their facilities are demonstrated. Subjects are informed how many calls will be comprised in this visit. On subsequent visits the subjects are merely informed that the procedure will be the same as before, with possibly a different number of calls. An example of some operational details required by an experiment can be found in 2.5.8.2, Table 4/2.5, of the *Handbook on Telephonometry*.

A.4.4 Data collection

Speech levels, and related data such as durations and activity factors, may be derived from the tape recordings, but are now normally measured on-line, by computer-controlled meters, and stored directly into computer files for subsequent analysis.

Two subjective responses are collected per conversation per subject by the experimenter. The essential data consists of the conversation-opinion score and the Difficulty decision.

These responses can be collected by using the primitive method of each subject recording his/her result on paper. An example of such a form can be found in 2.5.8.2, Table 5/2.5, of the *Handbook on Telephonometry*. As subjects, especially the younger generation, use keyboards quite extensively then collection of votes is acceptable using this method.

A.4.5 Treatment of results

This is a very extensive subject, and only a brief outline can be given here.

Each conversation gives rise to two conversation opinions on the scale: Excellent – Good – Fair – Poor – Bad (scored respectively 5, 4, 3, 2, 1), two votes on the Difficulty scale (scored 1 = Yes, 0 = No), two measured active speech levels and one value of duration. In particular cases information may be collected about other variables also; for example, video recordings may be made in order to observe how subjects hold their handsets, or other data may be derived from the opinion forms or from the audio recordings.

Calculation of separate standard deviations for each condition is not recommended. Confidence limits should be evaluated and significance tests performed by conventional analysis-of-variance techniques.

The usual assumptions underlying the analysis of variance are sufficiently nearly satisfied in the case of opinion score, active speech level and most other variables of interest; but they are not satisfied – particularly the assumption of constant residual variance – in the case of a binary variant like Difficulty score. In spite of this, experience confirms the observation made in other fields [11] that the analysis of variance technique is robust enough to give reasonable results even with such extreme departures from the statistically ideal conditions. The results from the first stage of the analysis of variance of the Difficulty scores should be regarded with some reserve; but once it is established that there are no unexplained abnormalities in the results and no unexplained conflicts with the outcome of the corresponding analysis of the Y_C results, then the second stage (detailed analysis of the averages from the End-Condition combinations) can be confidently undertaken with the aid of a mathematical transformation.

Detailed descriptions of the analysis can be found in the *Handbook on Telephonometry*, 2.5.9.

As a further aid to the review of the data, graphs if appropriate should be plotted showing the mean opinion score as a function of the parameter under test, e.g. Y_C versus circuit attenuation. On the graph the vertical axis should always be Y_C.

Annex B

Listening tests – Absolute Category Rating (ACR)

(This annex forms an integral part of this Recommendation)

B.1 Source recordings

In order to eliminate unwanted variability in the speech source, samples of speech having the desired standardized properties should first be prepared in recorded or stored form, as follows.

B.1.1 Recording environment

The talker should be seated in a quiet room with volume between 30 and 120 m^3 and a reverbation time less than 500 ms (preferably in the range 200-300 ms). The room noise level must be below 30 dBA with no dominant peaks in the spectrum.

The room noise characteristic should be reported in as complete a form as possible e.g. dBA, long-term spectrum, and amplitude-time distribution. It is desirable to record a 30-second sample of the room noise for detailed investigation if this proves to be necessary.

B.1.2 Sending system

Whatever sending system is chosen, e.g. local telephone or Intermediate Reference System (IRS) as specified in Recommendation P.48, the system should be calibrated according to the relevant Recommendation (e.g. Recommendation P.64), and the sending sensitivity-frequency characteristic should be reported in full.

It is recommended that the sending sensitivity characteristic of the connection is measured at least twice; at the beginning and end of the experiment. Any significant variation in the two measurements, when compared with each other, must be assessed by the experimenter as it may cast doubt on the validity of the experiment.

B.1.3 Recording system

The recording system must be of high (studio) quality and can be any of the following:

a) A conventional two-track tape recorder. The type of equalization must be stated, but IEC is recommended. High grade tape (low print-through, low noise) should be used at all times.

b) A two-channel digital audio processor with a high quality video cassette recorder or Digital Audio Tape (DAT) machine.

c) A computer-controlled digital storage system.

The third system is the best and most versatile, but practical reasons often dictate the choice of one of the other systems. In these, one of the two tracks should be used for recording the speech and the other for inserting control signals at a level and a frequency chosen to avoid crosstalk problems.

B.1.4 Speech material

The speech material should consist of simple, meaningful, short sentences, chosen at random as being easy to understand (from current non-technical literature or newspapers, for example), and made up into lists in random order, in such a way that there is no obvious connection of meaning between one sentence and the next. Very short and very long sentences should be avoided, the aim being that each sentence when spoken should fit into a time slot of about 4 seconds, occupying about half this time for the actual speech. Examples of sentences are shown in Table B.1.

The experimenter must decide how many sentences are required in each group to constitute a speech sample. A minimum of two and a maximum of five are recommended. The time interval between sentences, during which circuit noise may be heard and adaptive processes settle into new states, is also important. It is advisable to record the longest groups that may be needed, as it is always possible to obtain shorter groups by copying or replaying parts of longer ones.

Groups are combined into lists consisting of five or ten groups each, so that a complete list can be used as a series of samples subjected to the same treatment but with listening level or some other parameter varied when the list is reproduced.

TABLE B.1/P.80

Examples of speech material

You will have to be very quiet.
There was nothing to be seen.
They worshipped wooden idols.
I want a minute with the inspector.
Did he need any money?

B.1.5 Recording procedure

The following recording scheme has been extensively used and is recommended.

Speech is recorded from a linear microphone and low-noise amplifier with a flat frequency response as specified in IEC Publication 581-5. The microphone is positioned between 140 mm and 200 mm from the talker's lips. A windscreen is used if breath puffs from the talker are noticed.

The same speech is also recorded simultaneously from the sending output of an Intermediate Reference System (see Recommendation P.48), with the handset held in the normal manner, or other telephone if the investigation in view specifically requires it.

Two separate recording systems are used simultaneously: one for recording the wideband speech in one channel, and the other for recording the telephone speech in the corresponding channel. The other channel of each recording system is used for recording control signals as explained in 2.5 of the *Handbook on Telephonometry*.

This dual recording system ensures that the same speech is recorded in two forms (telephone speech and wideband speech). Normally only one of these is required in any one experiment, but there are occasions when it is necessary to use both, and it is an advantage in any case to be able to make comparative measurements on the two versions.

The active speech level, as defined in Recommendation P.56, is observed during recording. Care is taken during the recording process that the active speech level in both recording systems is between 20 and 30 dB below the overload point of the recording system for each sentence measured separately. Any group of sentences for which this does not hold is re-recorded.

It is recommended that the ratio of the active speech level to psophometrically weighted noise level (for definition see 3.2.3/P.83), SNR(p), on the recording media should be > 40 dB with an objective of 50 dB.

All speech samples used in one experiment should be different: this is essential for Listening-Effort tests and desirable for other types.

B.1.6 Talkers

There must be as many talkers as required in the design of the experiment (see B.3).

Talkers should pronounce the sentences fluently but not dramatically, and have no speech deficiencies such as stutter; they should adopt a speaking level that is comfortable to them as individuals and which they can maintain fairly constant.

B.1.7 Speech levels

The recordings when completed are played back, and the active speech level of each sentence is measured with a meter conforming to Recommendation P.56. The lists (announcements, sentences and control tones) are then re-recorded on to a second system with the necessary gain adjustments, so as to bring each group of sentences to the standardized active speech level specified below, and still preserve the proper time relationships between the sentences and the tone signals in the other channel.

For the narrow-band speech, the standardized level is derived by measuring and adjusting the narrow-band recorded signal directly; the recommended target is –26 dB (+ 0.5 dB) relative to the peak overload level of the recording system. The calibration tone has its r.m.s. level equal to the mean active level of the re-recorded speech.

For the wideband speech, account must be taken of the intended use of the recordings. It is sometimes appropriate to adopt the same levels as for telephone speech, but if the recording is intended for playback through a loudspeaker or artificial mouth, then the individual target speech levels should be such that equality is maintained at the output of the whole electric-acoustic replay chain.

B.1.8 Calibration signal

At the beginning of each resulting recording, 20 seconds or more of tone are inserted at the re-recording stage (for calibration purposes) at a level that is in a known relationship to the mean active speech level (most conveniently, equal to it). This calibration tone is normally at 1000 Hz, but may be at some other frequency if the recordings are intended for playing through systems (such as certain types of sub-band coder) that respond to 1000 Hz in a typical manner.

This tone can then be used later to adjust the mean input speech levels (see B.4.3).

B.2 Selection of circuit conditions

B.2.1 Speech input and listening levels

In the selection of circuit conditions particular attention should be applied to:

- range of input levels;
- range of listening levels:
  - there is no universal optimum listening level;
  - a variety of listening levels will occur in practice;
  - comparability considerations;
  - interactions may occur.

A detailed explanation of these aspects can be found in the *Handbook on Telephonometry*, 2.5.8.1.

B.2.2 Talkers

Since sophisticated processes often affect male and female voices differently, the experimental design should provide for two types of voice as a balanced factor; scores for male and female speech should be evaluated separately, only to be averaged if they yield main effects and interactions that are not statistically different.

Moreover, to reduce the danger that the results may depend heavily on peculiarities of the voices chosen, it is essential for more than one male and more than one female voice to be used in a balanced design.

B.2.3 Reference conditions

Every experiment should include reference conditions so that experiments made in different laboratories or at a different time in the same laboratory can be sensibly compared. Such reference conditions will depend on what is being assessed. For a digital system the reference conditions may include the Modulated Noise Reference Unit (MNRU) conforming to Recommendation P.81; other controlled degradations are appropriate in other cases (e.g. signal-to-noise ratio, see 3.2.3/P.83).

B.2.4 Other conditions

Besides the requirements of B.2.1 to B.2.3 inclusive, other conditions will be included depending on the purpose of the test. For instance, room noise might be a variable as well as bit-error rate for a digital system or Rayleigh fading for a radio system.

B.3 Design of experiment

The design of the experiment uses the same principles as given in A.2.

In addition, the experiment design must cater for the following:

a) Requirements of B.2;

b) Order-of-presentation effect.

For a given sample of subjects the test is limited in size by the maximum length of session possible without fatigue. If the experiment is too large to be catered for in one session then it is prudent to sub-divide into two or more sessions. Ideally no session should last for more than 20 minutes and in no case should a session exceed 45 minutes.

B.4 Listening test procedure

B.4.1 Listening environment

The listening room should meet the same conditions as the recording room (see B.1.1) with the exception that the environmental noise (see A.1.1.2.2) should be set to the appropriate level. See A.1.1.2.2.1 and A.1.1.2.2.2 for examples of noise spectra.

It is recommended that the noise level and spectrum are measured at least twice; at the beginning and end of the experiment. Any significant variation in the two measurements, when compared with each other, must be assessed by the experimenter as it may cast doubt on the validity of the experiment.

B.4.2 Listening system

Whatever listening system is chosen (e.g. local telephone system, Intermediate Reference System as specified in Recommendation P.48, a loudspeaker system), the system should be calibrated according to the relevant Recommendation (e.g. Recommendation P.64) and the receiving sensitivity/frequency characteristic should be reported in full.

It is recommended that the receiving sensitivity/frequency characteristic of the connection is measured at least twice; at the beginning and end of the experiment. Any significant variation in the two measurements, when compared with each other, must be assessed by the experimenter as it may cast doubt on the validity of the experiment.

B.4.3 Listening level

The gain of the system should be set in such a way that the calibration tone (see B.1.8) played back from the processed tapes produces the required listening level.

Variations in listening level, as required by the experiment design, can either be accommodated by use of

a) attenuators/amplifiers in the listening system; or

b) included in the processing or re-recording stage.

The second method is not recommended because it is difficult to maintain a high enough signal-to-noise ratio at low levels, and because flexibility and variety in the randomization are greatly reduced.

The listening level should always be recorded and information on this subject can be found in 2.5 of the *Handbook on Telephonometry*.

B.4.4 Listeners

Subjects taking part in listening tests are chosen at random from the normal telephone using population, with the provisos that:

a) they have not been directly involved in work connected with assessment of the performance of telephone circuits, or related work such as speech coding;

b) they have not participated in any subjective test whatever for at least the previous six months, and not in any listening-opinion test for at least one year; and

c) they have never heard the same sentence lists before.

If the available population is unduly restricted, then allowance must be made for this fact in drawing conclusions from the results.

In some cases screening of subjects may be necessary and a method based on Annex B/P.78 may be applicable.

B.4.5 Opinion scales recommended by the CCITT

Various five-point category-judgement scales may be used for different purposes. The layout and wording of opinion scales, as seen by subjects in experiments, is very important, and should follow the standard arrived at through years of experience. The following opinion scales are those most frequently used for CCITT applications and equivalent wording should be used depending on language which might result in small variations to the original English text:

a) *Listening-quality scale*

Quality of the speech	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The quantity evaluated from the scores (mean listening-quality opinion score) is represented by the symbol Y_{LQ} but where suffix notation is not available, the symbol Ylq is used.

b) *Listening-effort scale*

The heading of the listening effort opinion scale is particularly important. Without it, the other descriptions are liable to be seriously misunderstood.

Effort required to understand the meanings of sentences	*Score*
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

The quantity evaluated from the scores (mean listening-effort opinion score) is represented by the symbol Y_{LE} but where suffix notation is not available, the symbol Yle is used.

c) *Loudness-preference scale*

Loudness preference	*Score*
Much louder than preferred	5
Louder than preferred	4
Preferred	3
Quieter than preferred	2
Much quieter than preferred	1

The quantity evaluated from the scores (mean loudness-preference opinion score) is represented by the symbol Y_{LP} but where suffix notation is not available, the symbol Ylp is used.

NOTE – Examples of alternative subjective scales, which should only be used if the above three opinion scales do not meet the needs of the experimenter are given in 2.6 of the *Handbook on Telephonometry* and CCIR Report 751, Volume VIII.3, 1986.

B.4.6 Instructions to subjects

An example of typical instructions is given in Table B.2. The instructions must be given (verbally as well, if necessary) prior to commencement of the experiment. When the subject has understood the instructions, he/she should listen to the preliminary list and give his opinions. No suggestion should be made to the subjects that the preliminary samples include the best or worst in the range to be covered, or exhaust the range of conditions they can expect to hear. After the preliminary list, there should be sufficient time allowed for answering possible questions by the subjects. Questions about procedure or about the meaning of the instructions should be answered, but any technical questions must be met with the response, “We cannot tell you anything about that until the experiment is finished”.

TABLE B.2/P.80

Instructions to subjects

LISTENING EXPERIMENT No. ...

In this experiment you will be listening to short groups of sentences via the telephone handset, and giving your opinion of the speech you hear.

On the table in front of you is a box with five illuminated press buttons. When all the lamps go on, you will here .... sentences. Listen to these, and when the lamps go out, press the appropriate button to indicate your opinion on the following scale.

EFFORT REQUIRED TO UNDERSTAND THE MEANINGS OF SENTENCES

5	Complete relaxation possible; no effort required.
4	Attention necessary; no appreciable effort required.
3	Moderate effort required.
2	Considerable effort required.
1	No meaning understood with any feasible effort.

The button you have pressed will light up for a short time. Then the lamp will go out, and there will be a brief pause before all the lamps go on again for the next group of .... sentences.

There will be a longer pause after every ...... groups (each calling for an opinion). There will be a total of .... groups in this visit, and a similar number in your subsequent visit(s).

Thank you for your help in this experiment.

B.4.7 Statistical analysis and reporting of results

The numerical mean (over subjects) should be calculated for each condition at each listening level, and these means listed for initial inspection (so that effects such as those due to male and female speech can be seen).

Calculation of separate standard deviations for each condition is not recommended. Confidence limits should be evaluated and significance tests performed by conventional analysis-of-variance techniques.

NOTE – In the past, the equivalences, for example, Excellent = 4, Good = 3, Fair = 2, Poor = 1, Bad = 0 have often been used. Anyone using results from earlier experiments must be aware that the mean scores must all be increased by one to be comparable with those now obtained; otherwise there is no difference in the numerical processing that may be applied.

The method of analysis of the opinion scales of B.4.5.1 follows the principles stated in A.4.5.

As a further aid to the review of the data, graphs if appropriate should be plotted showing the mean opinion score as a function of the parameter under test, e.g. Y_{LE} versus circuit attenuation. On the graph the vertical axis should always be the Y_{LE} score.

Averaging the scores of the male and female talkers should be made with care and does not imply that this step would be warranted for a detailed study and interpretation of results unless the significance tests justify it.

Annex C

Quantal-Response Detectability Tests

(This annex forms an integral part of this Recommendation)

The best method for obtaining information on the detectability or some analogous property of a sound (such as echo) as a function of some objective quantity (such as listening level), is a quantal-response method similar in principle to that described in [12] using the "Three-operator without hidden loss method" for determining Reference and Relative Equivalents – this procedure is described in the *Handbook on Telephonometry*, 2.2. The main difference is that the subject's response is not a decision in the form "Reference" or "Test" (the designation of the louder of the two circuits), but a vote on a scale such as [13]:

Detectability opinion scale

A Objectionable

B Detectable

C Not detectable

where B is understood to mean "Detectable but not objectionable".

Scales of this sort, usually with three points, may be used in a variety of quantal-response tests; for example the scale as shown above may be used where the stimulus is echo, reverberation, sidetone, voice-switching mutilation, or interfering tones, while crosstalk and perhaps echo in some circumstances may be judged on the scale Intelligible – Detectable – Not detectable.

It is sometimes permissible to regard these votes as opinion scores, with values 2, 1, 0 respectively, and treat them in the same sort of way as listening or conversation opinion scores. But this is often unsatisfactory because the decisions on such a scale as detectability (see above) are not really equivalents of responses on a continuous scale – as votes on such scales as "Loudness preference" (see B.4.5.1), may be legitimately taken to be – but effectively embody two distinct dichotomies (for example detectable/not detectable and objectionable/not objectionable), which though not independent may nevertheless call different psychological processes into action: in other words, Objectionability or Intelligibility differs in kind, not merely in degree, from Detectability. For this reason a more profitable method of analysis is to express the probability of response according to each dichotomy separately, as a function of some objective variable, by fitting probit or logit equations, and then using the quantiles or other parameters as a basis of comparison between circuit conditions, in a manner analogous to that used in applying articulation scores.

The actual conduct of experiments of this type resembles that of listening-effort tests (see Annex B), but there are some differences. In particular it is advisable that the first presentation of the signal in each run should be at a high listening level, so that the listener is left in no doubt what kind of signal is a candidate for his decisions. Where sidetone or echo is involved, the subject will be required to talk as well as listen.

Simple audiometric measurements, as described in Recommendation P.78 [14], are usually performed on subjects who participate in these experiments, so that results can be expressed relative to their threshold of hearing.

For examples of the application of these techniques, see [15], 3.5.3 and 4.5.1.

Noise, fading and other disturbances are sometimes investigated by means of responses on a scale with many more points; for example [16],

A *Inaudible* – Noise completely undetectable

B *Just audible* – Noise can just be detected by listening carefully

C *Slight* – Noise detectable, but not disturbing

D *Moderate* – Noise slightly disturbing

E *Rather loud* – Noise causes appreciable disturbance

F *Loud* – Noise very disturbing, but call would be continued

G *Intolerable* – Noise so loud that the call would be abandoned, or operator asked to change the line.

These scales are more nearly of the quantized-continuum type, like the Loudness preference scale, and can be treated similarly.

Annex D

Degradation Category Rating (DCR) method

(This annex forms an integral part of this Recommendation)

D.1 Introduction

The Absolute Category Rating (ACR) test tends to lead to low sensitivity in distinguishing among good quality circuits. A modified version of the ACR (see Annex B) procedure, called the Degradation Category Rating (DCR) [17] procedure affords higher sensitivity. This procedure is adapted from the CCIR Recommendation [18] for evaluation of good quality circuits. The DCR procedure, which in particular uses an annoyance scale and a quality reference before each configuration to be evaluated, seems to be suitable for evaluating good quality speech.

D.2 Degradation Category Rating (DCR) procedure

D.2.1 Speech samples

Each configuration is evaluated by means of judgements upon four talkers reading two different samples. Each sample should comprise two sentences separated by at least one second. These two samples (S1, S2), hence four different sentences, should be selected from a wider corpus composed of phonetically balanced sentences so that the mean score obtained in evaluating reference (e.g. MNRU for digital processes) circuits for these sentences is about the same as that obtained for the wider corpus. Therefore the corpus consists of eight samples defined as follows:

- talker T1 reading samples S1, S2;
- talker T2 reading samples S1, S2;
- talker T3 reading samples S1, S2;
- talker T4 reading samples S1, S2.

This results in a repetition of the two samples during the test. But it is felt that this is not so critical for the procedure where a degradation is evaluated with regard to the reference, especially for good telephone quality where the intelligibility of speech is nearly perfect. The use of different samples for each configuration as is done in ACR experiments, where the speaker and the sentence effects are confused, could be one of the reasons for this procedure's lack of sensitivity. Some variations of this basic scheme are allowed: increase the number of talkers, mix sentence and talker effects, but each configuration must be evaluated on the same corpus.

D.2.2 Reference conditions

Reference conditions shall be included e.g. for digital processes multiplicative noise with Q values within the range 10 to 30 dB with a minimum of four steps is desirable.

A quality reference should be chosen to be inserted before each judgement. Usually source conditions are used, i.e. samples with no more degradation than those introduced by sending systems and limitations of frequency bandwidth. Thus, the choice of the quality reference depends on the application, i.e. for standard telephony, the source signal is 3.4 kHz bandwidth limited, for wideband telephony it is 7 kHz band limited and for high quality sound, the signal is 15 or 20 kHz band limited.

D.2.3 Stimulus presentation

The stimuli are presented to listeners by pairs (A-B) or repeated pairs (A-B-A-B) where A is the quality reference sample and B the same sample processed by the system under evaluation. The purpose of the reference sample is to anchor each judgement of the listeners. Some "null pairs" (A-A), at least one for each talker, are included to check the quality of anchoring. Using a reference and subjective judgements with respect to that reference is quite a common procedure in psychoacoustics. It tends to result in a good sensitivity for the overall evaluation by listeners. Samples A and B should be separated by 0.5 s and in a repeated pair procedure (A-B-A-B) the separation between the two pairs should be 1 s.

It seems that the classical order effect observed in a one-sample listening test (ACR for example) is not observed with the DCR procedure. Thus, only one random order of presentation can be used. Therefore the basic test and reference conditions will be eight times (four talkers × two samples) the number of nominal conditions. The timing for the response of listeners is the same as for the ACR test, i.e. 5 s between each presentation (pair or repeated pairs).

D.2.4 Test instructions

The subjects should be instructed to rate the conditions according to the five point degradation category scale as follows:

5 – Degradation is inaudible

4 – Degradation is audible but not annoying

3 – Degradation is slightly annoying

2 – Degradation is annoying

1 – Degradation is very annoying.

The quantity evaluated from the scores (mean degradation category rating score) is represented by the symbol Y_{DCR} but where suffix notation is not available, the symbol Ydcr is used.

D.3 Statistical analysis

Sensitivities can be quantified by means of a statistical multiple comparison test. When an *a posteriori* comparison of circuits is needed a Tuckey [19] honestly significant difference (HSD) test can be applied effectively. The HSD test is designed to make all pair-wise comparisons among the means and to determine the significance of the differences in the mean values.

Annex E

The threshold method for comparison of transmission systems with a reference system

(This annex forms an integral part of this Recommendation)

E.1 Introduction

By direct comparison of a transmission system with a reference system it is possible to assess the performance of the system under test in terms of a degradation characteristic of the reference system which can be varied and set to defined values. An example of such a characteristic is signal-to-noise ratio (for definition see 3.2.3/P.83), SNR(p). The method described here leads to a threshold of equality defined as 50% preference level between the MNRU and the digital system.

E.2 Testing procedure

A listening-only test procedure is used. A signal pair consisting of a reference signal and a test signal is presented to listeners, who are then asked to indicate which of the signals in the pair they judge to have the highest quality (preference rating). Subjective equivalence is defined as the reference value corresponding to the intersection point of the regression curve of the preference scores at the 50% preference level. An example of equivalent SNR obtained with hypothetical preference scores is shown in Figure E.1.

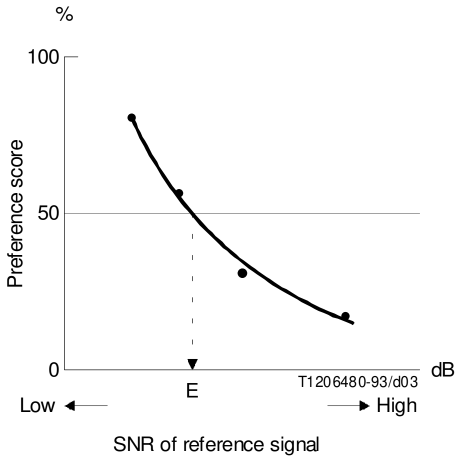


FIGURE E.1/P.80

Example of an equivalence threshold E with hypothetical preference scores

E.3 Presentation of signals

Reference signal A and test signal B are arranged in an equal number of A-B pairs and B-A pairs, and presented in random order. Several degradation levels spaced for example, at 2 dB intervals, are introduced in the reference path so that the range of preference scores extends from 20% to 80%, where the 50% preference lies in the middle of the degradation range. A timing diagram of the presentation is shown in Figure E.2.

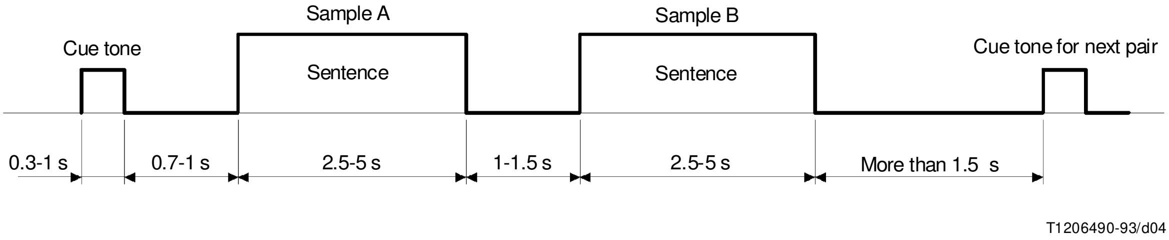


FIGURE E.2/P.80

Timing diagram of the presentation

The subject is required to make a judgement and respond by saying "A is better" or "B is better" (forced choice). The response "A equals B", or "No difference" is forbidden. The duration of the presentation should be limited to about six minutes in order not to tire the listeners. More listening samples may be presented after a suitable rest period. At least two, preferably four or five replications (repetitions of identical presentations) are recommended.

NOTE – If the reference system is available in hardware and the degradation characteristic can easily be changed between presentations, a simplified procedure can be used. In this case the balancing to equally perceived quality is done by the subject. The adjustment is made during the pause between the pairs. The reference signal is always presented first. Presentation continues until the subject reports that the equality threshold has been reached.

E.4 Speech sources

It is necessary to use short sentences spoken by at least two males and two females, preferably four or six of each; different sentences are required for each speaker. The duration should be 2.5-5 seconds for speech and less than 10-15 seconds for music signals. Clicks at the beginning and end of the samples must be avoided. A linear microphone of sufficient bandwidth should be used to record the source signals in a sound-absorbent room having an ambient noise of less than 20 dBA and a reverberation time of less than 0.3 seconds in the band 125-8000 Hz. If digital recording equipment is used, the quantizing noise level should be less than the noise level in 14-bit linear PCM.

E.5 Listening environment

A high-fidelity sound reproduction system should be used for the listening test. When listening is carried out with loudspeakers, the reproduction equipment should be studio-quality and the listening room should conform to CCIR Report 797 or IEC 268-13. If headphones are used, diotic (binaural) listening is preferable. The bandwidth shall be at least as wide as that of the system under test.

E.6 Listeners

Although it is preferred that listeners should be selected according to the description in the ACR method (see Annex B), this is not a strict condition in the pair comparison test. If the purpose of the listening test is to obtain the opinions of untrained listeners, untrained subjects are necessary. However, if this is not the purpose of the test, then trained listeners can be used and the reliability of the listening test can be extended by increasing the number of replications for each listener. The minimum number of listeners is six, but should preferably be twelve or more. Several subjects may listen simultaneously but it must be ensured that their responses are obtained independently.

E.7 Reliability

Since variations in preference score in subjective tests are assumed to conform to a t-distribution, the score variation width r which yields 95% reliability at score u ($0 \leq u \leq 1$) over the number (n) of trials (i.e. the number of repetitions for each presentation pair multiplied by the number of subjects and number of source signals) is presented in equation (E-1).

$$r = \pm\ t(n - 1,\ 0{,}05) \cdot \sqrt{u(1 - u)/(n - 1)} \tag{E-1}$$

NOTE – The threshold method is expected to give stable and reliable results even for high quality systems with little degradation.

Degradation can be introduced in the reference system e.g. by addition of white noise. For digital systems, multiplicative noise as defined in Recommendation P.81 (MNRU) is recommended. For wideband digital encoders, the use of a wideband MNRU as described in Annex A/P.81 is recommended. For some purposes shaped noise instead of white may be appropriate.

References

[1] VOIERS (W.D): Evaluating processed speech using the Diagnostic Rhyme Test, Speech Technology, Volume 1, No. 4, pages 30 to 39, January-February 1983.

[2] CCITT *The SIBYL method of subjective testing, Red Book*, Volume V, Supplement No. 5 to Recommendation P.74.

[3] BERANEK (L.L.): Noise and Vibration Control, pages 564 to 566, McGraw-Hill, 1971.

[4] ISO *Assessment of noise with respect to community response*, ISO Recommendation 1996, (1971).

[5] HOTH (D.F.): Room noise spectra at subscribers' telephone locations, *J.A.S.A.*, Vol. 12, pp. 99-504, April 1941.

[6] IEC Recommendation Publication 179, *Precision sound level meters*, 1965.

[7] CCITT Question 24/XII, Contribution COM XII-No. 120, *Noise inside light motor vehicles*, Study Period 1981-1984.

[8] CCITT Question 24/XII, Contribution COM XII-No. 134, *Internal vehicle noise spectra*, Study Period 1981-1984.

[9] CCITT Contribution COM XII-No. 208 *Comparison of the results of vehicle noise submitted by France and BT*, Study Period 1981-1984.

[10] STEVENS (S.S.): Psychophysics: Introduction to its perceptual, neural, and social prospects, *John Wiley and Sons*, 1975.

[11] CLARINGBOLD (P.J.): The within-animal bioassay with quantal responses, *Journal of the Royal Statistical Society,* Series B, Volume 18, No. 1, pages 133 to 137, 1956.

[12] CCITT *Measurement of Reference and Relative Equivalents*, Recommendation P.72, § 2.1.2, *Yellow Book*, Volume V, 1981.

[13] RICHARDS (D.L.): Telecommunication by speech, § 3.5.2, *Butterworths*, London, 1973.

[14] CCITT *Subjective testing method for determination of Loudness Ratings in accordance with Recommendation P.76*, Recommendation P.78, Annex B, *Blue Book*, Volume V, 1989.

[15] RICHARDS (D.L.): Telecommunication by speech, § 3.5.3 and § 4.5.1, *Butterworths*, London, 1973.

[16] Ibid, § 4.2.1.6.

[17] COMBESCURE (P) et al: Quality evaluation of speech coded at 32 kbit/s by means of degradation category ratings, Proc. ICASSP 82 (International Conference on Acoustics, Speech and Signal Processing), Vol. 2, Paris, May 1982.

[18] CCIR *Subjective assessment of the quality of television pictures (EBU)*, Document 11/17, Study Period 1978-1982.

[19] TUCKEY (J.W.): The problem of multiple comparisons, *Ditton*, Princetown University, Ed 1953.

Bibliography

GABRIELSSON (A): Statistical treatment of data from listening tests on sound-reproducing systems, Report TA No. 92, November 1979, *KTH Karolinska Institutet*, Department of Technical Audiology, S-10044 Stockholm, Sweden.

IEC Publication 268-13, Annex 3, § 3.3 (a condensed version of the above report).