INTERNATIONAL TELECOMMUNICATION UNION

ITU-T P.833

(02/2001)

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Methods for objective and subjective assessment of quality

Methodology for derivation of equipment impairment factors from subjective listening-only tests

ITU-T Recommendation P.833

(Formerly CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Subscribers' lines and sets	Series	P.30 P.300
Transmission standards	Series	P.40
Objective measuring apparatus	Series	P.50 P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of quality	**Series**	**P.80** **P.800**
Audiovisual quality in multimedia services	Series	P.900

For further details, please refer to the list of ITU-T Recommendations.

ITU-T Recommendation P.833

Methodology for derivation of equipment impairment factors from subjective listening-only tests

Summary

This Recommendation gives a methodology for deriving equipment impairment factors (*Ie*s) for digital signal processing devices, namely low bit-rate codecs with or without transmission errors.

The *Ie* values are derived from the results of subjective listening-only tests. They are intended to be used as an input to the E-model (see ITU-T G.107).

Source

ITU-T Recommendation P.833 was prepared by ITU-T Study Group 12 (2001-2004) and approved under the WTSA Resolution 1 procedure on 23 February 2001.

Keywords

E-model, equipment impairment factor, impairment factor method, low bit-rate codecs.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. 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 Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

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

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

ITU-T Recommendation P.833

Methodology for derivation of equipment impairment factors from subjective listening-only tests

1 Scope

This Recommendation describes the methodology for deriving equipment impairment factors (*Ie*s) from subjective listening-only tests. It is intended that it primarily be applied to determining *Ie*s for digital signal processing devices used in the network that have not otherwise been covered by the E-model. The equipment impairment factors derived by this methodology are intended to be used in the E-model (see ITU-T G.107). They will reflect the auditory impairments of the corresponding equipment in a listening-only mode.

2 References

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

- ITU-T G.107 (2000), *The E-Model, a computational model for use in transmission planning*.
- ITU-T G.108.1 (2000), *Guidance for assessing conversational speech transmission quality effects not covered by the E-model*.
- ITU-T G.113 (2001), *Transmission impairments due to speech processing*.
- ITU-T G.711 (1988), *Pulse code modulation (PCM) of voice frequencies*.
- ITU-T G.726 (1990), *40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)*.
- ITU-T G.727 (1990), *5-, 4-, 3- and 2-bit/sample embedded adaptive differential pulse code modulation (ADPCM)*.
- ITU-T G.728 (1992), *Coding of speech at 16 kbit/s using low-delay code excited linear prediction*.
- ITU-T G.729 (1996), *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)*.
- ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.
- ITU-T P.810 (1996), *Modulated noise reference unit (MNRU)*.
- ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.
- ITU-T P.831 (1998), *Subjective performance evaluation of network echo cancellers*.
- ITU-T Handbook, *Handbook on Telephonometry*, 1992.

3 Introduction

The equipment impairment factor method (ITU-T G.113) is now the only one recommended by ITU-T for describing the subjective effects of digital processes other than pure PCM on the integral quality for transmission planning purposes. It is based on the principle that individual transmission impairments can be transformed into "psychological factors", and that these psychological factors are additive on a "psychological scale". It is assumed that the transmission rating scale underlying the E-model (ITU-T G.107) has the properties of such a "psychological scale".

In transmission planning of modern telecommunication networks, it is important to consider the combined effects of different types of impairments occurring simultaneously in a connection. This is performed by computational models like the E-model. Individual sources of degradations, such as a non-optimum overall loudness rating or sidetone path attenuation, talker echo attenuation and delay, quantizing distortion, absolute delay, etc., are transformed into so-called impairment factors. Degradations due to low bit-rate codecs are taken into account by means of an equipment impairment factor. Whereas interrelationships exist between the other impairment factors due to the underlying instrumentally measurable quantities of the connection (e.g. loudness ratings), the equipment impairment factor is by its very definition independent of all the other impairment factors. It is only dependent on the digital process whose perceptual characteristics it aims to model. The final transmission rating for the entire connection is calculated from all the individual impairment factors, by subtracting them from a basic signal-to-noise ratio.

For asynchronous tandems of multiple codecs of the same type, or of multiple codecs of different types, it is assumed that individual equipment impairment factors are additive. The overall equipment impairment factor for the chain of codecs is then calculated as a simple sum of all the individual *Ie*s. Experimental test data collected provide evidence that this simple additivity is not satisfied for all the potential combinations of codecs [5]. In some cases order effects exist, i.e. the tandem of codec A followed by codec B results in a different degradation than codec B followed by codec A. These deviations from the pure additivity property will be examined in more detail.

Equipment impairment factors have been introduced as a simplified measure of the perceptual effects of non-waveform codecs on the integral transmission quality from mouth to ear, for narrow-band (300-3400 Hz) handset telephony. They are in no way an exact description of the effects related to each individual codec or codec tandem, which may be very diverse in their perceptual nature. Instead, they represent the relative degradation in comparison to other impairments occurring in a connection. In order to provide guidance on the quantitative amount of impairment introduced by such codecs, a framework of equipment impairment factors for the most common non-waveform codecs has been derived by ITU-T, see Appendix I/G.113. Derivation is based on many subjective listening-only tests carried out in different test laboratories, so as to guarantee stable values for all the codecs under investigation so far, as well as their relations to each other. If new equipment impairment factor values for different codecs have to be derived, then the overall consistency with the established framework is of primary importance if results are to be obtained that are valid for network planning. The methodology described in this Recommendation was designed to fulfil this requirement.

As long as no instrumental methods can be defined, equipment impairment factors have to be derived from subjective test data. Such tests are generally carried out in a listening-only mode, which allows several, different connections to be tested within one test session. As a result, equipment impairment factors derived in this way will only reflect the influence on the one-way speech transmission quality, and will disregard degradations in the conversational quality. It has not yet been proven that the quantitative degradation in the listening-only and in the conversational modes are similar, but for reasons of simplicity a strong interrelation is assumed. Users of transmission rating models, however, should be aware that differences may exist, and that most of the data derived only reflects the listening-only situation.

Subjective tests, even if carefully designed and carried out under controlled conditions, cannot provide quality ratings which are 100% reproducible under the same conditions. The composition and experience of the test panel, choice of test conditions and stimulus material, test set-up and environment lead to an inherent variability. This variability can also be found in the mean ratings calculated over a large number of individual responses. As a consequence, equipment impairment factors derived from one test will vary to a certain extent if compared to other test data. The degree of variability, however, should be limited to what can be tolerated in the transmission planning of networks, the latter including a reasonable safety margin. The comparison and integration of test data obtained under different experimental conditions is one way to reduce variability. The strict application of the experimental design and test method described in the following clauses is an additional one.

4 Selection of experiment parameters

In order to obtain valid and reliable results, some general requirements for the design of listening-only tests should be satisfied. Many of these requirements are described in detail in ITU-T P.800 and ITU-T P.830; therefore, wherever possible, reference to these Recommendations is made. Practical implications on how to carry out subjective tests can be found in the ITU-T Handbook on Telephonometry. As a general fact, listening-only tests do not achieve the same standard of realism regarding the actual service conditions experienced by telephone customers as conversation tests do. However, wherever possible, the most realistic approach should also be chosen in listening-only tests, in order to guarantee that results obtained in the relatively artificial laboratory situation have significance for the actual service.

4.1 Preparation of test material

The preparation of source recordings should satisfy – with the exception of the points explicitly raised hereafter – the description given in B.1/P.800, and in clause 7/P.830, for narrow-band systems. This rule applies to the recording environment, the sending as well as the recording system, and the recording procedure.

ITU-T P.800 and ITU-T P.830 recommend the use of simple, meaningful, short sentences as language material. Although tests carried out with this material will provide reasonable results, test realism will be enhanced by the use of longer, more diverse and meaningful material, such as short, meaningful test passages. Using such material, the listener's attention will be focused more on the contents and not uniquely on the form of the speech sample, an effect which corresponds better to actual service conditions. Useful language material can be obtained e.g. from non-technical literature, newspapers, etc. The overall length of the test passages will be similar to the maximum obtained for the isolated sentences, i.e. around 12-15 seconds. If transmission error conditions are included in the test, a sufficient exposure to the degraded speech parts should be ensured due to the length of the test material. The same holds true for background noise conditions, where a sufficiently long exposure to the environmental noise is required in order to obtain valid ratings.

The recommendations regarding the number of talkers given in 8.1.3.1/P.830 apply. If only the minimum number of two male and two female talkers is used, test results should be interpreted in this light, especially when differences to test data obtained elsewhere are detected.

Because digital speech codecs may be sensitive to the level of the input speech, the new codec under investigation should be tested at different levels relative to the overload point of the digital system. The recommendations given in 8.1.1/P.830 apply. In order to reduce the size of the experiment, reference conditions may only be tested at the optimum input speech level. For all test conditions, the listening level should be held constant at the preferred value. This is important because degradations due to a non-optimal *OLR* are already taken into account by the *Iolr* impairment factor of the E-model. This effect in the test data for deriving equipment impairment factors would therefore lead to unnecessarily pessimistic overall estimations.

The same considerations apply to ambient room noise at the send side (parameter *Ps* of the E-model). This effect is already taken into account by the E-model's basic signal-to-noise ratio *Ro*, and should consequently not be investigated in tests aiming at deriving equipment impairment factors.

NOTE – If strong background noise effects on the codec performance are to be expected, these codecs should be tested both with and without the application of ambient room noise at the send side, and at various noise levels (e.g. *Ps* ranging from 35 dB(A) to 75 dB(A)) and potentially different types of background noise (e.g. Hoth noise, speech babble noise, car environment noise, etc.). The recommendations of 8.1.8/P.830 should then be respected, with the exception of Note 1 in that clause – it is strongly advisable to perform source recordings in the noisy environment, because the effects of ambient noise on the talking behaviour proved to be of some significance (see the update of the E-model for this aspect). The outcome of the experiment is that a comparison can be made between the two cases: which includes both effects (codec and room noise) in one single equipment impairment factor for the codec under room noise conditions (then applying the default value for *Ps* in the E-model); or which caters for both effects independently via an *Ie* for the codec under quiet send side noise conditions and an inclusion of the effects related to *Ps* via *Ro* in the E-model. Potential differences between the two methods should be highlighted and should either result in a worst-case calculation when applied to network planning, or in applying a table of equipment impairment factors for the codec and for each room noise level condition.

If codecs are to be tested under conditions of transmission errors, the guidelines given in 8.1.4/P.830 should be respected. The outcome of such a test will be an integral equipment impairment factor for the specified codec and transmission error condition. This value should then be used in the E-model to cater for these joint effects. The derivation of computational formulae regarding impairment factors for transmission error rates for a whole codec class is at present being investigated by ITU-T Study Group 12.

Subjective test conditions will include both single codecs as well as codec tandemings. Single codecs are likely to be configured according to the transcoding scheme given in Figure 4/P.830. The quantizing effects of both A-/μ-law encoders and decoders have already been taken into account by the E-model impairment factor for speech signal synchronous impairments (*Is* calculated from *Iq*), and should ideally be disregarded in tests intended to derive equipment impairment factors *Ie*. The slight quantizing degradation, however, is not expected to have a measurable influence on the subjective test results. Configurations, as given in Figure 4/P.830, can therefore be used for single codec conditions. The expected additivity of equipment impairment for multiple codecs refers to asynchronous tandem conditions. Thus, multiple codecs should be configured, as described in 8.1.6.1/P.830. This refers to both multiple tandems of the same codec and to mixtures of several different codec types.

4.2 Selection of reference conditions

For reasons given in clause 3, subjective tests aiming at deriving equipment impairment factors should include a number of reference conditions. These references are necessary for anchoring impairment factor values, and they will guarantee that new equipment impairment factor values fit into the existing system given in Appendix I/G.113.

The choice of reference conditions is influenced by the test item, i.e. the specific codec under investigation. As a general rule, reference conditions should be similar in perceptual nature to the test items, so that the rating is based on the same underlying perceptual components. For example, a non-waveform codec, which provokes a metallic sound of the transmitted voice, should not uniquely be tested against a reference which introduces signal-correlated quantizing noise. As a consequence, the Modulated Noise Reference Unit, MNRU (ITU-T P.810) is in most cases not an adequate reference condition for low bit-rate non-waveform codecs. Until now there has been no recognized reference unit that produced a scalable distortion similar in perceptual nature to what has been found to exist for such codecs. It is therefore recommended that different types of waveform and non-waveform codecs be used as a reference for experiments deriving equipment impairment

factors. The exact reference conditions to be used vary depending on whether transmission errors are to be considered, and whether an additivity check has to be performed.

4.2.1 Reference conditions for low bit-rate codecs without transmission errors

When equipment impairment factors for non-waveform codecs disregarding transmission errors are determined, the set of 14 reference codec conditions given in Table 1 should be included in the subjective test conditions. This list has been chosen from well-investigated codecs to cover the whole range of *Ie* values and degradation types.

Table 1/P.833 – Reference conditions for low bit-rate codecs without transmission errors

No.	Abbreviation	Codec type	Reference	Operating rate (kbit/s)	*Ie* value
1	G.711	log. PCM	ITU-T G.711	64	0
2	GSM-EFR	ACELP	GSM 06.60, Enhanced Full Rate	12.2	5
3	G.726(32)	ADPCM	ITU-T G.721 (1988), ITU-T G.726, ITU-T G.727	32	7
4	G.728	LD-CELP	ITU-T G.728	16	7
5	G.729	CS-ACELP	ITU-T G.729	8	10
6	G.726(32) *x2*	ADPCM	ITU-T G.721 (1988), ITU-T G.726, ITU-T G.727	32	14
7	G.728 *x2*	LD-CELP	ITU-T G.728	16	14
8	GSM-FR *alt.* IS-54	RPE-LTP *alt.* VSELP	GSM 06.10, Full Rate, *alt.* IS-54	13 *alt.* 8	20
9	G.729 *x2*	CS-ACELP	ITU-T G.729	8	20
10	GSM-HR *alt.* PDC	VSELP	GSM 06.20, Half Rate, *alt.* Japanese PDC	5.6 *alt.* 6.7	23 *alt.* 24
11	G.726(24)	ADPCM	ITU-T G.726, ITU-T G.727	24	25
12	G.729 *x3*	CS-ACELP	ITU-T G.729	8	30
13	GSM-FR *x2* *alt.* IS-54 *x2*	RPE-LTP *alt.* VSELP	GSM 06.10, Full Rate, *alt.* IS-54	13 *alt.* 8	40
14	G.726(16)	ADPCM	ITU-T G.726, ITU-T G.727	16	50

NOTE – *x2, x3* designates double respective triple asynchronous tandeming of the same codec.
alt. designates that either of the two codecs or codec tandems can be used for this test condition, resulting in either bit rate and/or impairment factor value.

It is important to check the additivity of the newly derived equipment impairment factor in the framework of other equipment impairment factor values defined so far. If such an additivity check is not performed, the property of a simple summation of equipment impairment factors in order to cater for codec tandems should not be regarded as valid. Table 2 gives a minimum number of ten additional reference conditions (Nos. 15-24) which should, in any case, be included in the test set to allow for a rough additivity check. It is preferable, however, to test inter-codec tandem operations with a larger set of similar conditions, including triple tandems in different codec orders.

Table 2/P.833 – Reference conditions for the additivity check in tandem operation of low-bit rate codecs without transmission errors

No.	Tandem operation	Reference codec type	Operating rate (kbit/s)	*Ie* value
15	G.726(32)*(new codec)	ADPCM	32	7 + *Ie*(new codec)
16	G.728*(new codec)	LD-CELP	16	7 + *Ie*(new codec)
17	G.729*(new codec)	CS-ACELP	8	10 + *Ie*(new codec)
18	GSM-FR*(new codec) *alt.* IS-54*(new codec)	RPE-LTP *alt.* VSELP	13 *alt.* 8	20 + *Ie*(new codec)
19	GSM-HR*(new codec) *alt.* PDC*(new codec)	VSELP	5.6 *alt.* 6.7	23 + *Ie*(new codec) *alt.* 24 + *Ie*(new codec)
20	(new codec)*G.726(32)	ADPCM	32	*Ie*(new codec) + 7
21	(new codec)*G.728	LD-CELP	16	*Ie*(new codec) + 7
22	(new codec)*G.729	CS-ACELP	8	*Ie*(new codec) + 10
23	(new codec)*GSM-FR *alt.* (new codec)*IS-54	RPE-LTP *alt.* VSELP	13 *alt.* 8	*Ie*(new codec) + 20
24	(new codec)*GSM-HR *alt.* (new codec)*PDC	VSELP	5.6 *alt.* 6.7	*Ie*(new codec) + 23 *alt.* *Ie*(new codec) + 24

NOTE – A*B designates asynchronous tandeming of codecs A and B, B followed by A.

alt. designates that either of the two codec tandems can be used for this test condition, resulting in either bit rate and/or impairment factor value.

4.2.2 Reference conditions for low bit-rate codecs with transmission errors

When equipment impairment factors for non-waveform codecs operating at low bit-rates under the effects of transmission errors are determined, the same reference conditions as given in Table 1 should be applied. This allows a relationship to be established between *Ie* values defined in the context of the E-model and those values actually measured in the specific experiment. In addition to these 14 conditions, a minimum number of $n = 10$ supplementary reference conditions including perceptively noticeable degradation due to transmission errors (random bit errors, random packet loss, bursty packet loss, or propagation errors in terms of error patterns as defined for the GSM codecs) should be applied.

These n supplementary reference conditions should be chosen from the already defined values given in Appendix I/G.113 for equipment impairment factors under transmission error conditions. They should cover the whole degradation range expected for the codec under investigation, as well as the type of transmission error applied to this codec. Especially if the codec under investigation makes use of some kind of error recovery strategy (e.g. in repeating previous packets or interpolating between adjacent frames), the reference conditions with transmission errors should possibly include codecs that apply similar strategies.

For codecs with transmission errors, reference conditions allowing for an additivity check should be included in the test set as well. Unfortunately, when different error rates are to be tested, this additivity check can lead to an experimental size that is barely manageable. For this reason, no mandatory list of reference conditions is given here. The considerations of clause 6, regarding the interpretation of *Ie*s that have been derived without performing an additivity check, apply.

4.3 Test environment

The listening environment should meet the requirements given in B.4.1/P.800. The noise level at the receive side should be checked to ensure that it satisfies the default parameter settings of the E-model, i.e. an A-weighted level of $Pr \leq 35$ dB(A), but also respects the limits of NC25 or NR25 set in ITU-T P.800. Ambient noise at receive side has already been taken into consideration by the E-model *Ro* parameter, and should consequently not be introduced in tests for determining *Ie*s.

Equipment impairment factors are to be used in conjunction with the E-model, which makes predictions for narrow-band (300-3400 Hz) handset telephony. The listening system should show the modified IRS receive characteristic described in Annex D/P.830. The calibration recommendations given in B.4.2/P.800 should be followed. As was argued in 4.1, the listening level should be kept constant at the optimum level, i.e. at 79 dB SPL at the ear reference plane.

4.4 Listening panel

The general requirements on the listening panel given in B.4.4/P.800 should be satisfied. A relatively large number of listening subjects are preferable in order to reduce inter-listener variance. However, provided there is a representative composition of the listening panel, a minimum number of 24 to 32 subjects will often suffice.

5 Test method

In principle, listening-only tests for the purpose of deriving equipment impairment factors should be conducted to provide absolute ratings, i.e. according to the "Absolute Category Rating" (ACR, see ITU-T P.800) method in cases where category scales are used, or with a corresponding method when continuous scales are used. Test subjects rate each stimulus individually on the scale provided for this purpose.

NOTE – Paired comparison test methods (e.g. "Comparison Category Rating", CCR) can be used to decide on the rank order of very similar equipment impairment factor values for different codecs. With the CCR method, a direct comparison of the codecs or codec tandems under investigation can be made (in contrast to the description in Annex E/P.800, both samples are then processed). However, such values first have to be derived from absolute ratings, and then in a second post-hoc test session the comparative rating can be obtained.

Subjective listening-only tests carried out for the purpose of deriving equipment impairment factors consist of two or three parts, depending on whether transmission errors are to be taken into account or not. These three parts reflect the five steps to be taken in the derivation methodology for impairment factors (see clause 6), and not necessarily the division in test sessions. Part A (steps 1 and 2) is always necessary, and consists of deriving a first value of *Ie* based on one single encoding and decoding process. An interpolation line is established which is necessary for all future transformations of actual test data to the framework of *Ie*s that has been defined so far. Part B (steps 3 and/or 5) consists of a check of the additivity property, both for pure tandems of the codec under investigation and mixed tandems with other codecs for which equipment impairment factors have already been defined. Part C (step 4) contains the additional steps to be taken when the codec is investigated under conditions of transmission errors.

Part B should always be carried out if additivity for the codec under investigation has not been proven. If the additivity does not seem to have been fulfilled, this fact should be highlighted when presenting the test results. Applying impairment factors for codecs that do not satisfy the additivity is questionable. As a minimum requirement, the application should be limited to telephone circuits where only one single coding and decoding process occurs. However, part B may be omitted if *Ie* values are to be determined for a codec with transmission errors for which an *Ie* value for the error-free case has already been defined.

5.1 Experiment design

The general requirements with respect to the experimental design given in A.2/P.800, should be fulfilled. A randomization of test stimuli according to the graeco-latin square type (cf. the ITU-T Handbook on Telephonometry) is preferable; however, due to the large number of test stimuli, such a design is relatively complicated. Other design strategies, as cited in ITU-T P.800, are therefore acceptable.

5.2 Choice of test stimuli

The choice of test stimuli will depend on whether transmission errors are taken into account. In any case, all 14 reference conditions given in 4.2.1 (Nos. 1-14) have to be included. As a minimum requirement, the codec under investigation should be included in single operation at three different input speech levels. For an additivity check, at least all 10 reference conditions (Nos. 15-24) defined in 4.2.1, plus the codec under investigation in double and triple asynchronous tandeming operation should be included in the test set. It is preferable to include further tandem conditions with well-investigated codecs (potentially triple tandems) in order to increase the stability of the results regarding a potential additivity of the derived equipment impairment factor. If possible, tandem conditions should be included symmetrically (i.e. codec A followed by codec B as well as codec B followed by codec A) in order to detect potential order effects. The additivity check may be omitted if *Ie* values for transmission error conditions are to be determined for a codec for which the *Ie* value in the error-free case has already been defined.

If the new codec is investigated under conditions of transmission errors, the range of transmission errors should cover the one that is likely to be encountered most frequently in network operation (*m* error conditions). The specific error rates chosen should be sufficiently finely graded so that the overall results remain stable when single subjective data points show the characteristics of outliers. In addition to the *m* conditions involving the codec under test, at least $n = 10$ additional reference conditions described in 4.2.2 have to be included in the test set. It is preferable to include more of such conditions if test design and resources allow.

Table 3 summarizes the test conditions to be included in the different parts of the experiment.

Table 3/P.833 – Overview of test conditions for the different parts of the experiment. Different test parts are described in clause 5

Part	Purpose	Test conditions	Mandatory/ Optional	Min. overall Σ test cond.
A	Determination of *Ie* for the new codec in error-free conditions	References 1-14	Mandatory	17
		New codec in single operation, at 3 speech input levels	Mandatory	
		Additional low-bit rate codec references	Optional	
B	Additivity check	References 15-24	Mandatory	12
		New codec alone in double and triple tandem operation	Mandatory	
		New codec in double and triple tandem operation with other codecs	Optional	

Table 3/P.833 – Overview of test conditions for the different parts of the experiment. Different test parts are described in clause 5 *(concluded)*

Part	Purpose	Test conditions	Mandatory/ Optional	Min. overall Σ test cond.
C	Determination of *Ie* for the new codec in transmission error conditions	Min. n = 10 references according to 4.2.2	Mandatory	$n + m$
		New codec in single operation in different transmission error conditions (m conditions)	Mandatory	
		Additional references according to 4.2.2	Optional	

5.3 Presentation method

In order to obtain more or less "absolute" ratings, test stimuli are presented to the listening subjects one by one. Subjects have to rate each stimulus individually on the scale provided for this purpose. The overall number of test stimuli is limited by the maximum session length that is possible without fatigue. Indications of B.3/P.800 apply. It is preferable, however, that all the reference conditions and the conditions including the codec under test be included in one test session, in order to achieve a better comparability of the results. If this is not possible (e.g. because more than the minimum number of speakers have been chosen) test sessions have to be split.

5.4 Opinion scales

The choice of an adequate measurement scale is important for the outcome of the subjective experiment. It is preferable to use a rating scale that is strongly related to the scale of the transmission rating factor *R* or to the (herewith linearly related) scale of impairment factors.

Unfortunately, there is no validated methodology available which would allow direct measurements on the impairment factor scale to be made. As a consequence, the traditional listening-quality ACR scale, as described in B.4.5/P.800 (MOS scale), is recommended. The use of the MOS scale requires a non-linear step in the result transformation procedure from the 5-point category scale to the scale of impairment factors (S-shaped transformation curve). This may lead to a loss in information. In Appendix I, information regarding a different scale (the so-called CR-10 scale) is provided which reflects the logarithmic behaviour of perceptual magnitudes, thus requiring only a linear transformation to the impairment factor scale. As the CR-10 scale has not yet been fully verified in the context of telephone impairment scaling, the MOS scale is currently the only recommended scale for deriving equipment impairment factors.

5.5 Instructions to test subjects

For every scaling experiment it is important that the rating task is absolutely clear to the test subjects. The wording of the heading and labels of the scale should be followed as precisely as possible, in the subjects' native language, which may result in small variations from the original English text. The general recommendations on instructions to test subjects given in B.4.6/P.800 should be fulfilled.

5.6 Analysis of results

The statistical evaluation of test data should satisfy the general requirements of experimental data analysis, as stated in B.4.7/P.800, as well as in the ITU-T Handbook on Telephonometry. Mean values of subjective ratings can be calculated first on a per speaker and per input level basis, for each circuit condition separately. If an analysis of variance reveals statistically significant differences due

to these factors, these should be highlighted in the description of test results, and caution should be exercised in their interpretation. If no statistically significant differences are found, ratings on individual stimuli can be mixed to form a mean rating per circuit condition.

6 Derivation of equipment impairment factors

The methodology for deriving equipment impairment factors from subjective listening-only tests carried out as described above consists of three to five steps, depending on whether transmission errors are taken into account or not:

Step 1: Scale transformation of the subjective test data.

Step 2: Derivation of a stable *Ie* value for the codec under test, in single codec operation without transmission errors, via a linear interpolation of the test results.

Step 3: Additivity check.

If transmission errors are under investigation, the following additional steps are to be taken:

Step 4: Derivation of stable *Ie* values for different transmission error conditions in single codec operation.

Step 5: Additivity check.

Step 3 may be omitted if the equipment impairment factor for the error-free case has already been defined, and only the corresponding *Ie* values for transmission errors are determined. In this case, the *Ie* value derived in step 2 should not significantly differ from the one already defined, so that the interpolation line accurately reflects the error-free case for the codec under investigation. In practical tests, step 5 can sometimes only be carried out roughly. It should be noted that if the amount of effort put into this step is limited, this will also limit the validity of the derived impairment factors with regard to tandem operation. This fact should clearly be stated in the description of test data analysis.

6.1 Necessary amount of data

As the principal outcome of data analysis, mean ratings (mean opinion scores, *MOS*) over all listeners, speakers and input levels will be available for the following test conditions:

- 14 reference conditions not involving the codec under investigation (reference conditions No. 1 to 14 in Table 1);
- the test condition involving the codec under investigation alone, calculated as a mean over three speech input levels;
- unless the additivity check may be omitted, two conditions involving the codec under investigation alone, in double and triple tandem operation;
- eventually further reference conditions not involving the codec under investigation;
- unless the additivity check may be omitted, ten conditions involving the codec under investigation and codecs for which *Ie* values are already known, i.e. mixed tandem conditions (reference conditions Nos. 15 to 24 in Table 2);
- eventually further mixed tandem conditions;
- in the case of investigation of transmission errors, a minimum of $n = 10$ reference conditions with transmission errors not involving the codec under investigation;
- in the case of investigation of transmission errors, all m conditions involving the codec under investigation with transmission errors;
- in the case of transmission errors, eventually further tandem conditions involving the codec under investigation with transmission errors in tandem operation with itself or other codecs.

If transmission errors are taken into account, each error condition (random bit error rate, burst error rate or error pattern) is regarded as a separate codec condition. No effort is made to derive conclusions regarding the dependence between codec performance and the error rate at this stage of data analysis.

6.2 Scale transformation (Step 1)

All subjective test results first have to be transformed on the scale of the equipment impairment factor *Ie*. The subsequent computations are then carried out on this scale alone.

Mean subjective ratings on the 5-point ACR listening-quality scale (*MOS*) require a transformation which is not available in closed formulae. *Ie,sub* values can be derived from the relation between *R* and *MOS* given in the E-model (see equation (B.4)/G.107), namely:

$$\begin{array}{rc} \text{for } MOS = 1.0: & R = 0 \\ \text{for } 1.0 < MOS < 4.5: & MOS = 1 + 0.035 \cdot R + R \cdot (R-60) \cdot (100-R) \cdot 7 \cdot 10^{-6} \\ \text{for } MOS \geq 4.5: & R = 100 \end{array} \quad (1)$$

This equation has to be resolved either numerically or graphically, using e.g. the plot given in Figure B.2/G.107. From the resulting values for *R*, the corresponding *Ie,sub* values can be calculated by defining the *R*-value for reference condition 1 (cf. Table 1) as an anchor, thus:

$$Ie,sub = R(\text{condition No.1}) - R(\text{test condition}) \quad (2)$$

This equation results in *Ie,sub* for the reference condition No. 1 always set to 0. It has to be noted that *Ie,sub* values derived in this way from *MOS* values may become negative if the corresponding *MOS* values are higher than the one for reference condition No. 1. This effect can be disregarded here if it only happens for one reference condition, because the subsequent linear interpolation may transform *Ie* values back to the positive part of the *Ie* scale.

The outcome of step 1 is an *Ie,sub* value for each test condition. It reflects the specific test condition, and it is not necessarily consistent with equipment impairment factors defined so far.

6.3 Linear interpolation of the test results (Step 2)

For all 15 reference conditions of Table 1, as well as possibly for all supplementary reference conditions involving only codecs for which *Ie* values have already been defined, pairs of expected equipment impairment factors *Ie,exp* and observed values *Ie,sub* are now available. These pairs can be represented as a scatter plot, see an example in Figure 1. A linear interpolation using a straight line

$$Ie,sub = a \cdot Ie,exp + b \quad (3)$$

can now be made. The coefficients *a* and *b* are determined numerically, approximating all the reference pairs in a least-squares sense. Alternatively, but with less precision, the approximation can also be made graphically on the scatter plot. Figure 1 shows the example of such an approximation and the corresponding coefficients *a* and *b*.

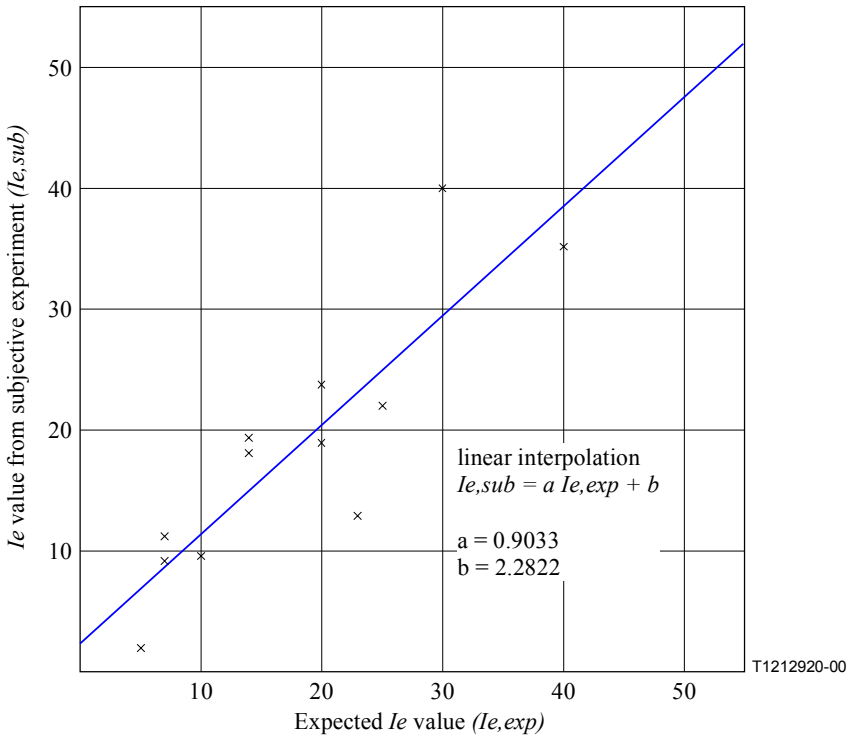


Figure 1/P.833 – Example of a linear interpolation of reference test data (values are taken from [5])

From this approximation, a stable equipment impairment factor value for the codec under test (*Ie,exp*) can be derived using equation (3) or its graphical representation. The *Ie,exp* value derived in this way will normally satisfy the framework of equipment impairment factors the interpolation line has been derived from, namely for the codecs included in reference conditions Nos. 1 to 14. However, it does not necessarily satisfy the additivity property underlying the impairment factor principle.

In rare cases the linear transformation may result in a negative *Ie* value for the codec under investigation. This might occur if the related subjective ratings are close to or better than the one for reference condition No. 1 of Table 1/G.711. In this case, *Ie* should be set to zero instead.

Because the *Ie,sub* value for the codec under test has been derived from three speech input level conditions, it is based on three times more data than the corresponding reference conditions, and should consequently be regarded as relatively stable. The linear interpolation of the reference codec conditions will deviate to a certain extent from the ideal diagonal line *Ie,sub* = *Ie,exp*, due to the characteristics of the specific experiment. This does not in general falsify the equipment impairment factor derived for the new codec, unless very large values for *b* (indicating that an experimental bias was present), or values for *a* significantly lower than 1.0 (indicating a misuse of the scale) are obtained. In the latter cases the source of errors should be investigated, and the experimenter has to decide whether reasonable equipment impairment factor values can be derived from the experiment.

6.4 Additivity check (Step 3)

The equipment impairment factor derived in step 2 does not necessarily satisfy the additivity property of *Ie*s. This has to be checked for both tandems of the new codec alone and mixed tandems with codecs for which *Ie* values have already been defined. Using the *Ie* value for the new codec derived in step 2, for all the tandems of the new codec alone, for all the reference conditions of Table 2, as well as for all further possible conditions involving mixed codecs, pairs of observed values *Ie,sub* and expected values *Ie,exp* are available. These pairs can be represented in the same scatter plot illustrated in Figure 1. All major deviations from the interpolation line should be noted and investigated, as they may question the applicability of the additivity of impairment factors.

If more than three out of twelve tandem conditions (two pure tandems of the codec under investigation and ten reference tandem conditions Nos. 15-24, see Table 2) show major deviations from the interpolation line, the additivity property should not be regarded as having been satisfied. In this case, the equipment impairment factor derived from the experiment will not properly represent the degradations occurring in tandem operations of the new codec.

6.5 Derivation of *Ie* values for transmission error conditions (Step 4)

For all *n* reference conditions with defined *Ie* values under transmission error conditions (see 4.2.2) pairs of subjectively determined equipment impairment factors *Ie,sub* and expected ones *Ie,exp* exist. These pairs can be added to the scatter plot derived in step 2. An investigation of the scatter plot may reveal:

a) that all data points group around the interpolation line found in step 2; or

b) that the new points group around a different interpolation line, with different ordinate crossing *b* and/or different slope *a*; or

c) that no grouping at all can be observed.

In case c), it is questionable whether meaningful equipment impairment factors for transmission error conditions can be derived. The test set-up and run should therefore be checked for any errors that might have occurred, and will consequently have to be repeated.

In case a), a new interpolation line can be calculated, this time from all the reference data points (reference conditions 1-14 and the additional *n* reference conditions defined in 4.2.2). Case b) reflects either a (different) bias for the transmission error conditions (different coefficient *b*) or a different application of the scale by the test subjects, perhaps due to a different underlying perceptual dimension (different slope *a*). If the coefficients of the interpolation line for the *n* transmission error conditions are meaningful (see step 2 for a discussion), then this different interpolation line should subsequently be used.

In both cases a) and b), the new interpolation line can be used to determine relatively stable *Ie,exp* values for the new codec under transmission error conditions, following the procedure of step 2. Graphs or tables can be drawn from these values and will simplify a plausibility check. The minimum consistency to be reached is to have non-decreasing equipment impairment factor values for increasing transmission error rates. Equipment impairment factor values derived in this way do not necessarily satisfy the additivity property. Moreover, they are based on relatively less data than the value derived for the error-free condition in step 2, because different speech input levels have been omitted. This allowance often has to be made for reasons of practicability, the number of test conditions being limited by the available resources. If major inconsistencies are detected, the experiment should be rerun, this time including more speech input levels and/or speakers to base a decision on.

6.6 Additivity check (Step 5)

A similar additivity check as in step 3 should be carried out on the equipment impairment factor values derived in step 4, using all the available tandem conditions of the codec under investigation with transmission errors introduced and other codecs for which impairment factors have already been derived. Unless this step can be carried out on sufficient data, the additivity of the derived impairment factors cannot be regarded as having been satisfied. The number of reference tandem conditions of Table 2 (10), each applied to a representative subset of the *m* transmission error conditions involving the new codec, can be regarded as sufficient in this case.

7 Interpretation of derived equipment impairment factor values

Due to the inherent variability of subjective ratings, equipment impairment factor values derived using this methodology cannot be expected to represent very exact quantitative measures of impairment that a subject in a specific situation would experience. In contrast, *Ie* values should be regarded as simplified values for network planning purposes only. *Ie* values will significantly differ between the experiments they have been derived from. The defined values of Appendix I/G.113 have mostly been derived using different sources of subjective test data, and can therefore be regarded as stable. Nevertheless, they have not been proven to be additive in all cases [5]. This property is therefore an item of further study.

Equipment impairment factors derived from subjective test data can only reflect the conditions under which they have been obtained. For the methodology described here, the listening-only test environment introduces a notable limitation. The *Ie* values will consequently only reflect degradations of the one-way speech transmission quality, and no specific conversational impacts. The latter are partly addressed by ITU-T G.108.1. However, as far as low bit-rate codecs are concerned, no fixed relationships between the impairment in the listening-only situation and the corresponding one in a conversational situation have been derived. Some investigations into this are described in [6], showing that such a potential relationship depends on the specific impairment as well as the quality dimension.

Another validity limitation has to be noted for *Ie* values under conditions of transmission errors. These errors are sometimes introduced with a specific frequency pattern (e.g. burst errors) or on specific speech frame units (e.g. packet loss). Unless it has been proven by means of auditory experiments using test subjects, it cannot be stated that the predictions obtained in this way will also be valid for other frame lengths or error distributions. Speech under transmission error conditions and the derivation of formulae for different error conditions are for further study.

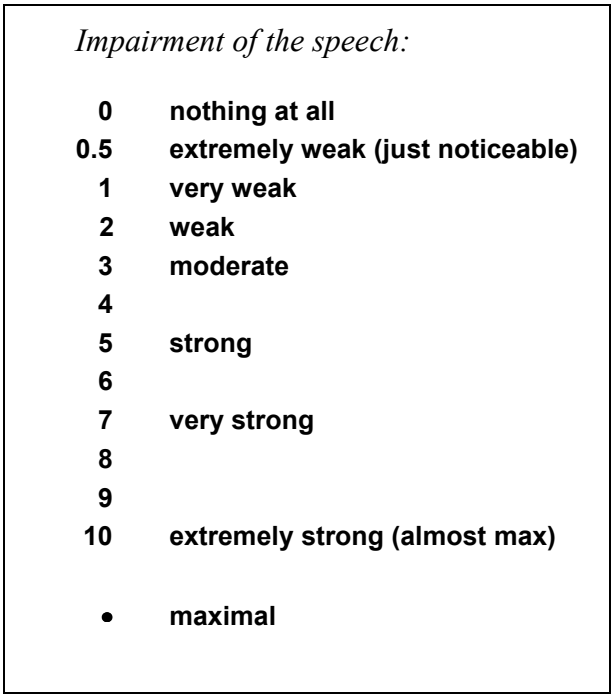
APPENDIX I

Rating scale related to impairment factors

For the determination of equipment impairment factors, a rating scale is desirable that is strongly related to the scale of the transmission rating factor *R* or to the (herewith linearly related) scale of impairment factors. Such a rating scale has been investigated by ITU-T Study Group 12 [6]. Unfortunately, it has not been fully verified for the given purpose. Therefore, the description of the scale and its application to derive equipment impairment factors is given for information only. The currently recommended scale in the context of this Recommendation is the MOS scale.

When considering the drawbacks of both category rating and magnitude estimation methods, category-ratio scaling has considerable advantages when it is to be used for the purpose of equipment impairment factor determination. It combines the advantages of the category scale (enabling absolute level determination) and the ratio scale (enabling the determination of relations between impairments, thus a characteristic of the transmission rating scale *R* or the impairment

factor scale). The easiest and most widely used version of such a scale is the CR-10 scale according to [4]. The positions of the verbal anchors of a category scale have been changed in order to reproduce the growth function that can be obtained by magnitude estimation. The CR-10 scale limits its numbers from 0, 0.5, 1, …10, but includes the possibility of using decimals or fractions or giving ratings above 10. It has been copyrighted by the author [4] in order to avoid misuse. A reproduction of the scale is given in Figure I.1, and copies are provided, together with a detailed description of its use, in [4].



NOTE – The scale is copyrighted.

Figure I.1/P.833 – CR-10 category ratio scale

For the CR-10 scale, two additional instructions should be given to the test subjects. The first is that the number 10 corresponds to the worst telephone communication quality ever experienced. This value is regarded as the main scale anchor, because it is assumed that perceptual intensities are approximately the same for different people at each individual's subjective maximum exertion. The second instruction is that test subjects should consider starting with the verbal expression first and should then choose the number that goes with it. Ratings can be provided as entire numbers, decimal numbers or fractions, and range between 0 and infinity. However, practical experience shows that numbers higher than 15 are rarely used by test subjects.

When using the CR-10 scale to derive equipment impairment factors, mean subjective ratings (*CR10mean*) can be linearly transformed into the equipment impairment factor scale using the relation

$$Ie,sub = 10 \cdot CR10mean - 5 \qquad \text{(I.1)}$$

The outcome is a raw equipment impairment factor value *Ie,sub* that can be regarded as the result of Step 1 of the derivation methodology, see 6.2.

APPENDIX II

Bibliography

[1] ETSI ETS 300 961 (1997), *Digital cellular telecommunications system; Full rate speech; Transcoding* (GSM 06.10).

[2] ETSI ETS 300 969 (1997), *Digital cellular telecommunications system; Half rate speech; Half rate speech transcoding* (GSM 06.20).

[3] ETSI ETS 300 726 (1997), *Digital cellular telecommunications system; Enhanced full rate (EFR) speech transcoding* (GSM 06.60).

[4] BORG (G.): Borg's Perceived Exertion and Pain Scales, *Human Kinetics Pub.*, Illinois, 1998.

[5] ITU-T Contribution COM 12-69 (1998), *E-model predictions and the impairment factor principle for low-bit-rate codecs and quantizing distortion: Analysis of test results.* Source: Federal Republic of Germany.

[6] MÖLLER (S.): Assessment and Prediction of Speech Quality in Telecommunications, *Kluwer Academic Publishers*, Boston, 2000.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communications
Series Y	Global information infrastructure and Internet protocol aspects
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2002