International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.804

(10/2017)

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

Methods for objective and subjective assessment of speech and video quality

Subjective diagnostic test method for conversational speech quality analysis

Recommendation ITU-T P.804

International Telecommunication Union

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
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	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 speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	**Series**	**P.800**
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

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

Recommendation ITU-T P.804

Subjective diagnostic test method for conversational speech quality analysis

Summary

Recommendation ITU-T P.804 describes a subjective methodology for assessing and diagnosing the quality of transmitted speech in a telephone conversation. In addition to a score for the overall conversation quality, the methodology yields overall quality scores for three perceivable phases in a telephone conversation: listening, speaking, and interaction, as well as scores for their corresponding seven perceptual dimensions. Four of the perceptual dimension scores represent degradation associated with the listening phase, two are associated with the speaking phase, and one is associated with the interaction phase. Each of the perceptual dimension scores are based on ratings of the amount of degradation present in one system condition. The method is designed to be used with naïve subjects. The dimension scores can be used to provide diagnostic information on the causes of system degradations. The method is meant as a complement to standard conversation tests.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.804	2017-10-29	12	11.1002/1000/13397

Keywords

Conversational speech quality evaluation, diagnostic evaluation of conversational speech quality, multi-dimensional quality assessment, subjective testing.

---

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). 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.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

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 not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.804

Subjective diagnostic test method for conversational speech quality analysis

1 Scope

This Recommendation describes a subjective test methodology which is able to assess and diagnose the quality of speech in a "telephone conversation" scenario. Common conversation tests, as described in [ITU-T P.800] and [ITU-T P.805], provide valid methods for the overall conversational quality, but do not give insights into reasons for possible quality losses. In addition, common conversational tests lack analytic ability, since naïve participants concentrate on the conversation flow. To circumvent these problems, this Recommendation describes a test methodology that specifically allows participants to perceive each phase of a conversation separately, in addition to a natural conversation, and yields overall conversational quality scores as well as quality scores for each phase (listening, speaking, interaction). In addition, scores for seven underlying perceptual dimensions of conversational speech quality are provided. These scores enable the analysis of conversational speech quality for diagnosis and optimization.

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. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[ITU-T P.805] Recommendation ITU-T P.805 (2007), *Subjective evaluation of conversational quality*.

[ITU-T P.806] Recommendation ITU-T P.806 (2014), *A subjective quality test methodology using multiple rating scales*.

[ITU-T P.835] Recommendation ITU-T P.835 (2003), *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

MDS Multidimensional Scaling

MOS Mean Opinion Score

PCA Principal Component Analysis

RNVT Random Number Verification Task

SCT Short Conversation Test

SD Semantic Differential

5 Conventions

None.

6 Introduction to conversational speech quality analysis

To provide diagnostic information about the quality of transmitted speech, ITU-T recommends using multiple rating scales in subjective experiments in [ITU-T P.806]. The approach targets assessing perpetual dimensions to give deeper insight into possible quality loss. The recommended method refers to the passive listening-only situation and gives additional information to overall quality mean opinion score (MOS) absolute category rating (ACR)-experiments, as recommended in [ITU-T P.800]. However, the listening-only situation, only partly agrees with reality in telecommunications. Thus, conversation tests to assess the quality in an interactive situation have been designed and described in [ITU-T P.805]. The recommended methods do not provide diagnostic information. For this, a set of seven perceptual dimensions for a conversational situation are proposed in [b-Köster2014]. The proposed dimensions cover the three possible phases/situations of a conversation: listening, speaking, and interaction. See [b-Guéguin2008].

To identify the relevant proposed dimensions, each phase has been analysed in detail, applying a (I) pairwise similarity experiment with a following multidimensional scaling (MDS) and a (II) a sematic differential (SD) experiment with a following principal component analysis (PCA).

Applying both methods in separate experiments in [b-Köster2014] and [b-Wältermann2010] resulted in the following set of perceptual dimensions for a conversational situation (see Table 1): the listening phase is comprised of four dimensions: *noisiness*, *discontinuity*, *coloration* and *loudness*; the speaking phase is comprised of two dimensions: *impact of one's own voice on speaking* and *degradation of one's own voice*; the interaction phase is comprised of only one dimension: *interactivity*.

Table 1 – Overview of the seven identified and proposed perceptual quality dimensions for a conversational situation

Conversational phase	Perceptual dimension	Description	Possible source
Listening phase	Noisiness	Background noise, circuit noise, coding noise	Coding, circuit or background noise
	Discontinuity	Isolated and non-stationary distortions	Packet loss
	Coloration	Frequency response distortions	Bandwidth limitations
	Loudness	Important for the overall quality and intelligibility	Attenuation
Speaking phase	Impact of one's own voice	How is the back-coupling of one's own voice perceived	Sidetone and echo
	Degradation of one's own voice	How is the back-coupling of one's own voice degraded	Frequency distortions of the sidetone and echo path
Interaction phase	Interactivity	Delayed and disrupted interaction	Delay

Since the proposed dimensions were identified in separate listening, speaking, and conversation tests, the dimensions were validated in a sophisticated conversational experiment in which each of the conversational phases, as well as all of the proposed perceptual dimensions were addressed [b-Köster2015a].

The outcome of the experiment showed that in traditional conversation scenarios the proposed dimensions are difficult to identify. Thus, this Recommendation describes a test method that specifically allows participants to perceive each phase separately, in addition to a natural conversation paradigm. In addition, the described method allows for directly quantifying the proposed seven dimensions within one single experiment. The method enables analysis of conversational speech quality for diagnosis and optimization and enables an increase in the number of conditions to be assessed.

7 Test methodology

7.1 Dimension rating scales

The subjective method provides a means for quantifying seven quality relevant perceptual dimensions in a conversational situation (noisiness, discontinuity, coloration, loudness, impact of one's own voice on speaking, degradation of one's own voice, and interactivity) directly by means of seven descriptive scales. In addition, the overall conversational quality and the overall quality for each individual phase are gathered.

Each dimension scale is dedicated to one particular dimension. The scales are labelled with the antonym-pairs describing the corresponding dimension. This enables for directly quantifying separate scores for each perceptual dimension present in a conversational situation. The overall rating scales and the graphical scale layout for the dimensions are shown in Figure 1. The continuous scales were chosen over traditional ACR scales because they showed to be more sensitive [b-Köster2015b]. While the labels on the left of the scales describe no impairment in the relating dimension, the labels on the right describe the maximum impairment. Thus the scales are considered to be unipolar. A detailed

description of usage and definition of the scales, as given to test participants, can be found in Annex A.

Overall Quality

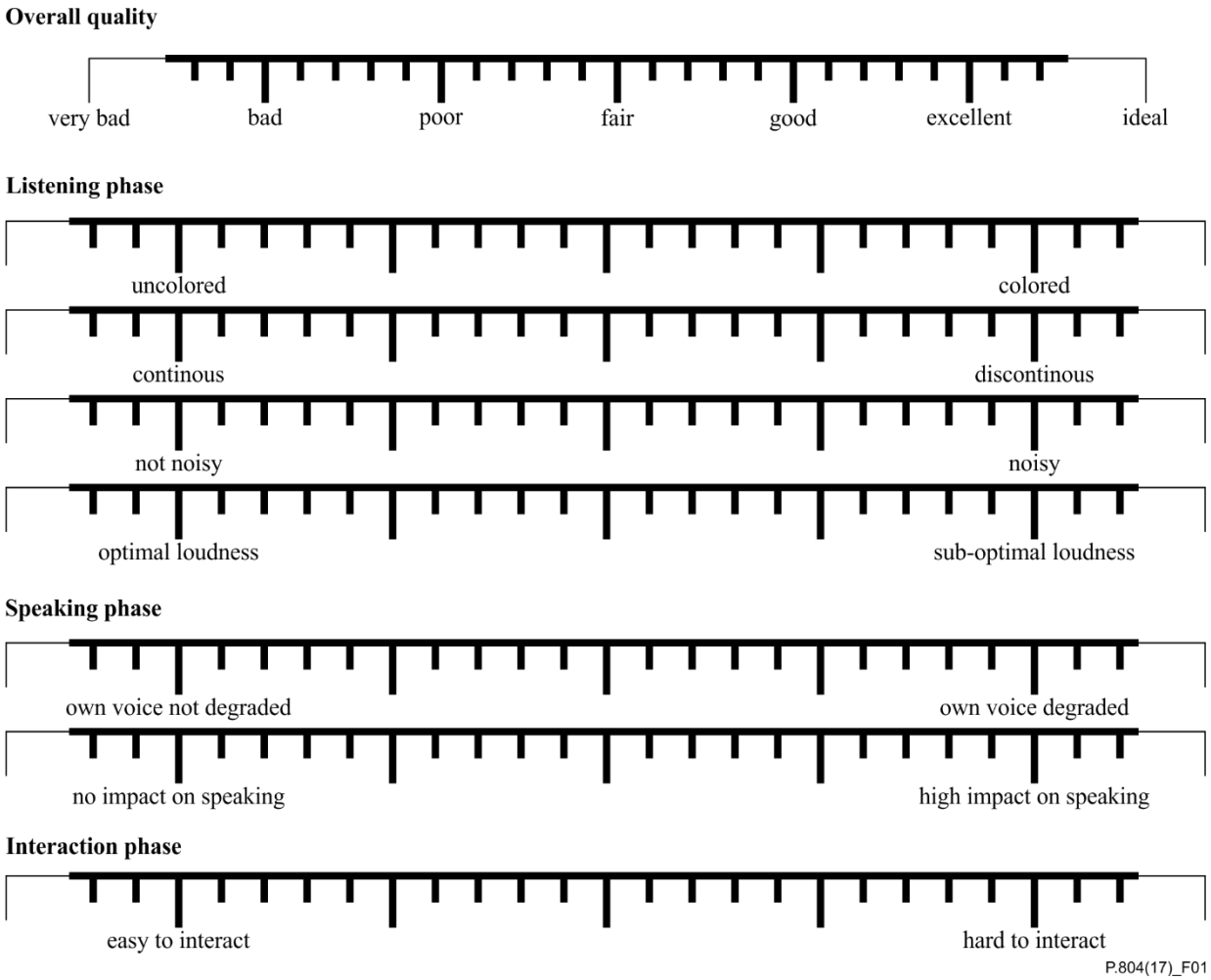


Figure 1 – Dimension scale design

7.2 Test design

The method follows common paradigms for subjective conversational tests as described in [ITU-T P.805]. For each condition, or transmission system properties under test, two participants in two separate rooms according to [ITU-T P.800] are required. The basic test setup can be seen in Figure 2.

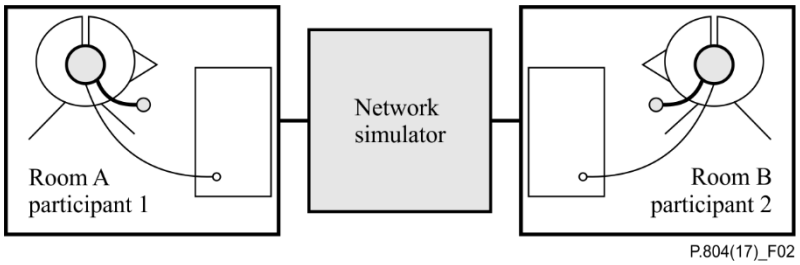


Figure 2 – Test method set-up

The test method specifically allows participants to perceive each phase separately, in addition to a natural conversation paradigm. Therefore, the test method to assess one condition is composed of three sessions:

1 In the first session, the task of the two participants is to conduct a short conversation test (SCT) scenario according to [ITU-T P.805]. The SCTs were used because their tasks represent everyday-life situations and provide a reasonable degree of interaction while being limited to an acceptable test duration. Thus, this session represents a regular everyday-life conversational scenario of about 2-4 minutes. After each scenario, the participants are asked to judge the overall quality (according to [ITU-T P.800] using the scale shown in Figure 1), and then the seven perceptual dimensions representing all phases of a conversation.

2 The second session addresses the listening and speaking phases. One of the participants is asked to read two sentences out loud while the other participant listens to what is being read. The sentences and procedures of the speaking part are similar to [b-Appel2002] and [b-Köster2014]. The listening part is analog to [b-Wältermann2010]. After the first sequence, the participants change roles so that each participant has to both speak and listen. For each sequence, the participants are asked to judge the overall quality of the speaking as well as the two dimensions for the speaking phase and the overall quality of the listening as well as the four dimensions for the listening phase.

   Examples for the speaking part can be found in Appendix I.

3 The third session addresses the interaction phase. This task is supposed to be sensitive for possible delays in the transmission system. Therefore, random number verification tasks (RNVTs) are used as per [ITU-T P.805]. The participants are asked to judge the overall quality of the interaction and the interactivity representing the interaction phase.

An overview of the test procedure for both participants can be seen in Figure 3.

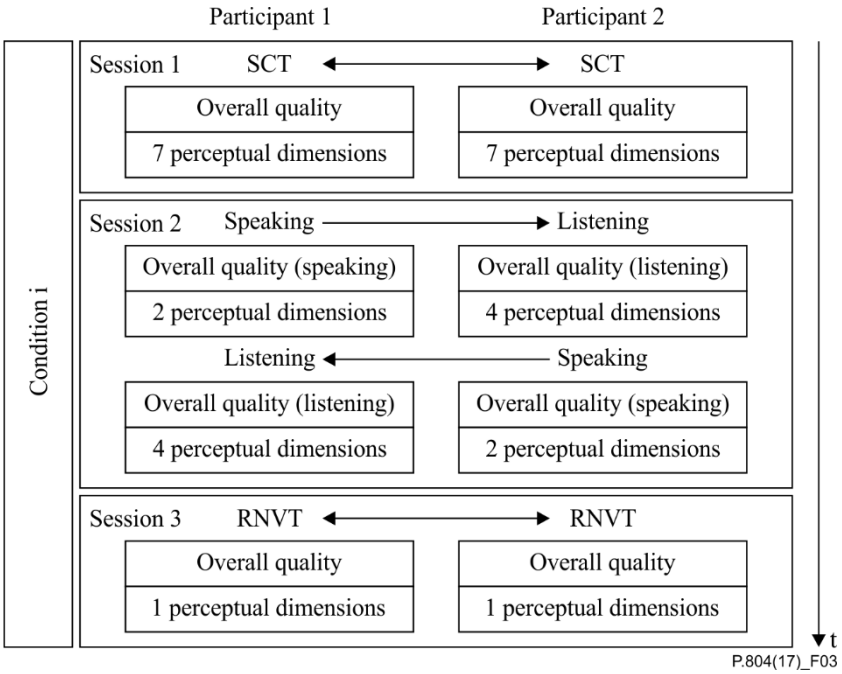


Figure 3 – Overview of the test procedure. SCT – short conversation scenario, RNVT – random number verification task

7.3 Dimension rating scheme

The dimension rating scheme of the test method is comparable to the scheme for analysing noisy signals [ITU-T P.835]. Each of the three separate sessions of the test method includes an assignment (speaking, listening, SCT, or RNVT) as well as an overall quality and a dimension assessment task.

As these assessment tasks follow the same procedure, only the rating task for session I will be described in detail.

After the overall quality assessment, according to [ITU-T P.800], the dimension scales (see Figure 1) are presented separately and consecutively. This is to reduce the bias due to the presentation order. Before participants are asked for their ratings, they are asked to conduct the given task once. Afterwards, the participants first give their judgments on the overall quality and second on the seven perceptual dimensions. The detailed rating schema for session I is shown in Figure 4.

The conditions to be assessed are presented in randomized order. Additionally, the order of the dimension scales is permuted for each participant. The schema is shown in Table 2. For each participant the order of the scales is held constant to avoid confusion of the scales.

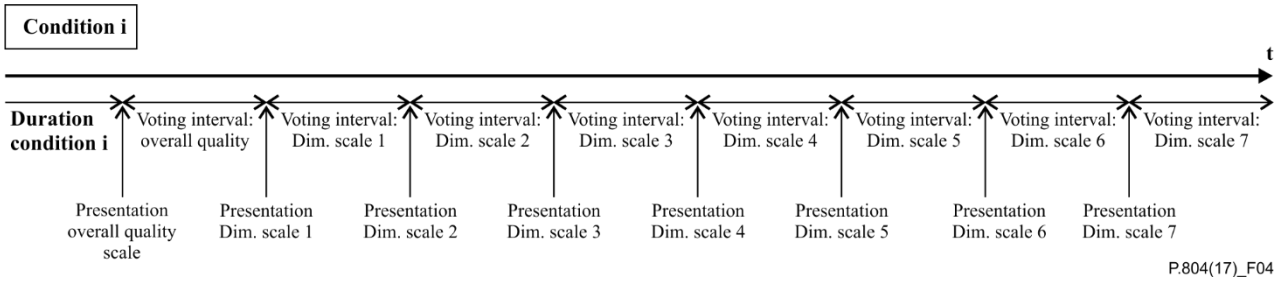


Figure 4 – Condition and scale presentation and rating for session 1

Table 2 – Presentation order of the dimensions scales. Noi – noisiness, Dis - discontinuity, Col – coloration, Lou – loudness, Ios – impact of one's own voice on speaking, Dos – degradation of one's own voice, and Int – interactivity

Participant	Dim scale 1	Dim scale 2	Dim scale 3	Dim scale 4	Dim scale 5	Dim scale 6	Dim scale 7
1	Dis	Col	Noi	Lou	Ios	Dos	Int
2	Col	Noi	Lou	Ios	Dos	Int	Dis
3	Noi	Lou	Ios	Dos	Int	Dis	Col
4	Lou	Ios	Dos	Int	Dis	Col	Noi
5	Ios	Dos	Int	Dis	Col	Noi	Lou
6	Dos	Int	Dis	Col	Noi	Lou	Ios
7	Int	Dis	Col	Noi	Lou	Ios	Dos
…	...	…	…	…	…	…	…

7.4 Instruction and training

A detailed written description of the test method is given to participants to ensure an equal level of knowledge (Annex A). First, the instruction gives an overview of the scales and how they should be used. It is explained, that in the experiment the characteristics of a conversation are supposed to be judged and that this judgment is done on seven scales. Each scale is labelled with an attribute at each end that describes the characteristic to be judged. The scales are described in detail using the highly correlated attributes according to the SD experiment conducted to extract the perceptual dimensions. Second, the test procedure (the three sessions and the relating tasks) is explained. These instructions can be found in Annex A.

In the training the two participants run through one test sequence as described in Figure 3. This is done to ensure that the participants get to know the test procedure and become familiar with using the scales and the test method.

In addition, optional training should be conducted to ensure that the test subjects get to know the test procedure as well as get familiarized with the usage of the scales and the test method. For this, the

test procedure as well as the ratings (overall quality and dimensions) should be practiced. Thus, one possible training exercise could look like the following:

The test subjects execute the test procedure (as described in Figure 3) twice. In the first run, the first session (SCT) is degraded with a condition related to the dimension noisiness. The second session (speaking and listening) is alternately degraded with conditions related to the dimensions discontinuity and impact of one's own voice. In the third session (RNVT), the test subjects will not be confronted with a degradation. In the second run, the first session is degraded with a condition related to the dimension Coloration. In the second session, the subjects will again be confronted with two alternating conditions, one related to the dimension loudness and one related to the dimension degradation of one's own voice. Finally, in the third session, the transmission system will introduce delay to trigger the dimension interactivity. The scheme of the training is illustrated in Table 3.

With this possible training, the two test subjects are introduced to the test procedure and the seven perceptual dimensions. In addition, the test subjects get to know the characteristics of all perceptual dimensions and practice using the rating scales. The training of the third session (no degradation vs. *interactivity*) is particularly useful to ensure the test subject's sensitivity for a transmission delay. As possible conditions, the test conditions (or adapted conditions) described in Appendix II could be used. However, the training is just a recommendation; it might also be possible to run through the test procedure only once to ensure that the test subjects understand their tasks.

Table 3 – Scheme of possible training for the proposed test method

	Session	**Perceptual dimensions Test subject 1**	**Perceptual dimensions Test subject 1**
Training run 1	1 (SCT)	Noisiness	Noisiness
	2 (listening/speaking)	Discontinuity	Impact of one's own voice
	2 (speaking/listening)	Impact of one's own voice	Discontinuity
	3 (RNVT)	none	none
Training run 2	1 (SCT)	Coloration	Coloration
	2 (listening/speaking)	Loudness	Degradation of one's own voice
	2 (speaking/listening)	Degradation of one's own voice	Loudness
	3 (RNVT)	Interactivity	Interactivity

Annex A

Test instructions – translated from German – Analysis of speech quality in a conversational situation

(This annex forms an integral part of this Recommendation.)

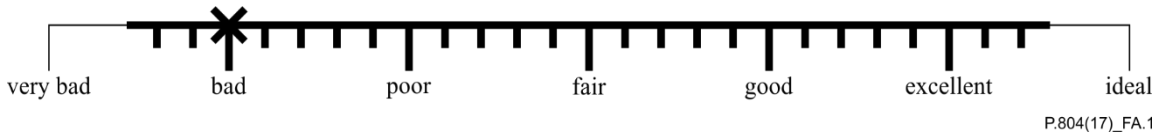
Thank you very much for taking part in this experiment! Please take the time to read the instructions thoroughly. Should you have any questions, please do not hesitate to ask the experiment supervisor.

You are taking part in a conversational experiment, in which the properties and the perception of a conversation are to be evaluated. To do this, you and your conversational partner will be put into a conversational situation in which various conditions will be presented in three phases. These phases are as follows: (1) a short conversational scenario phase, (2) a speaking and listening phase, and (3) an interaction phase. After each phase, you are to evaluate your perception and the properties of the phase. The exact procedure for individual phases will be described in detail later and will be made clear in a test run.

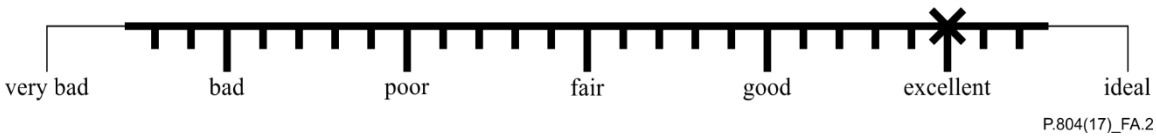
The characteristic of each condition is to be evaluated with the help of different scales. Please familiarize yourself with the scales and how to use them now.

Every scale has a term at each endpoint (e.g., not noisy/noisy). You are to evaluate to what extent the characteristics of a condition can be described by the terms on the scale. There are two different kinds of scales. The scale for overall quality has describing terms for every point, while the scales for the characteristics (e.g., not noisy/noisy) only have terms at their endpoints. The usage of these two scale types is analogous and is explained in detail for the overall quality.

The **overall quality** of a condition is to be evaluated. If you are of the opinion that the overall quality of the condition is bad, mark a cross at the position "bad" as shown next:

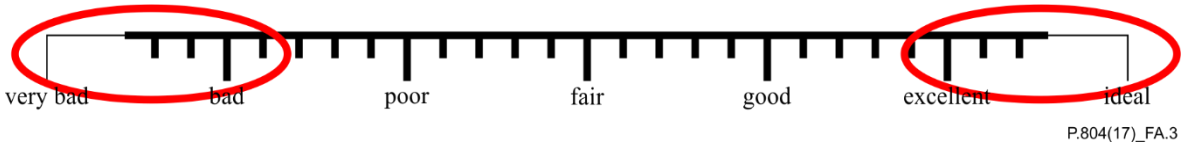


If you think that the overall quality of the condition is excellent, mark a cross at the position "excellent" as shown next:



For your evaluation, you can freely use the entire scale. The markings on the scale are there to provide points of reference to support your evaluation. You can even use the spaces in between the markings if you do not wish to settle for one of the markings.

And you can always use the "overflow areas", beyond the terms, if you feel the terms are not sufficient for your evaluation, as shown next:



The usage of the scales is analogous for the different characteristics. Overall there are seven characteristics to evaluate.

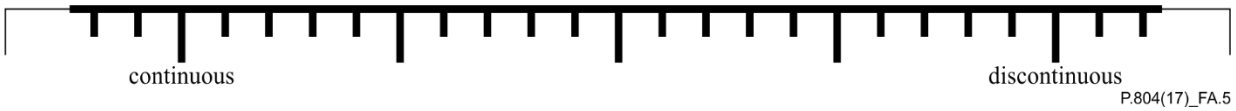
1 Noisiness

The first scale has the terms "not noisy" and "noisy" and appears as follows:

With this scale, you are to evaluate the **noisiness** of what you heard in a condition. The terms "not noisy" and "noisy" can be described with the terms *noisless* and *not hissing* and *noisy* and *hissing*, respectively.

2 Discontinuity

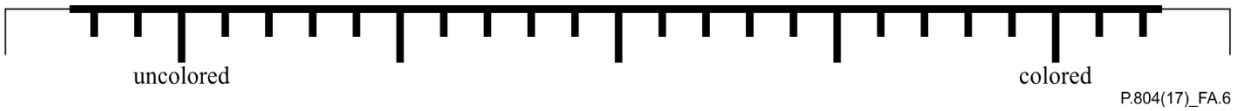
The **discontinuity** of what you heard in a condition is to be evaluated with the second scale:



The term "continuous" means that what you heard in a condition is completely *even*, *firm*, *not chopped* and *not frayed*. The term "discontinuous" can be described with terms like *uneven*, *wobbly*, *chopped* or *frayed*.

3 Coloration

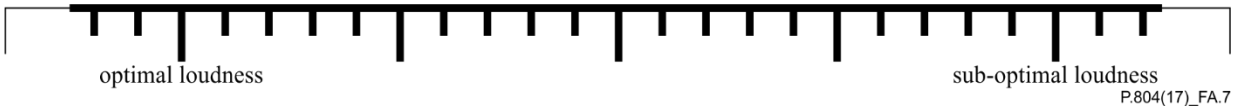
The **coloration** of what you heard in a condition is described by the third scale:



The term "uncolored" means that what you heard in a condition is *direct*, *close*, *full* and *not nasally*. The term "colored" means the heard sound is *indirect*, *far*, *thin* and *nasally*.

4 Loudness

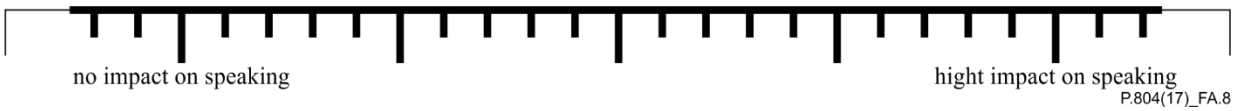
The fourth scale describes the **loudness** of the heard sound in a condition:



If the heard sound is neither too loud nor too quiet then the volume is optimal. If that is not the case then the volume is not optimal.

5 Impact of one's own voice

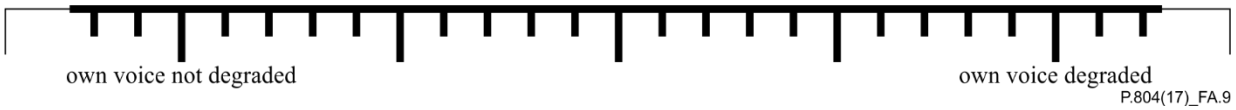
The **impact on your speaking through the hearing of your own voice** is to be evaluated with the fifth scale:



In some conditions, you will be confronted with back-coupling of your own voice. This means that you will hear your own voice. With this characteristic you are to evaluate if this does or doesn't negatively impact your speaking. More to the point, you are to evaluate if speaking while hearing yourself is *not distracting, not irritating, fluid* and *does not take concentration* or if it is *distracting, irritating, not fluid* and *takes concentration*.

6 Degradation of one's own voice

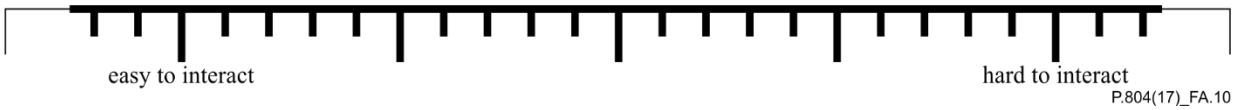
The sixth scale is for the evaluation of the **degradation of your own voice** while having back-coupling:



The term "own voice not degraded" means that you hear your own voice unaltered, e.g., you hear your voice *not distorted*, *echofree* and with *optimal volume*. The term "own voice degraded" can be described with terms like *distorted, echoed, thin* or *not optimal volume*.

7 Interaction

The seventh scale is used to evaluate the **interactivity** of a condition:



The term "easy to interact" means that the interaction between you and your conversational partner is easy. That means you have an *effective*, *pleasant*, *easy*, and *agile* interaction. The term "hard to interact" means that the interaction is *ineffective, unpleasant, hard* and *sluggish*.

For each scale, the following applies: Use as described for the scale for "overall quality" (see above).

Now that you have familiarized yourself with the meaning of the characteristics, the phases and the procedure of the test will be described.

<u>Phase 1 – Conversation</u>

Explanation

In this phase you and your partner simulate a telephone conversation. One of you will be the caller, who wants something and calls a company/organization/institution. The other one will be the company/organization/institution.

Explanation of symbols

As the caller, you have the following symbols:

This symbol means: You are the caller.
Please wait until the experiment supervisor asks you to start the first conversation.

Next to this symbol is the reason for your call.
E.g., I want to buy a ticket!

Next to this symbol are the conditions that should be incorporated into the exchange of information.

E.g., I want to buy a ticket! → *BUT cheaply if possible!*

Next to this symbol you are to record all information that you need from your telephone partner.

Next to this symbol is all of the information that your partner needs and that you should provide at some point during the conversation.

Next to this symbol is a question to which neither you nor your conversational partner have any information. You are to discuss this question briefly and come to a satisfactory conclusion together.

If you get called, you have the following symbols:

This symbol means: You are getting called. Wait until you hear the ringtone and then pick up.

Next to this symbol is information from which you are to sort out the information that your partner needs.

E.g., Prices for train tickets for adults, students, children, seniors, etc.

Next to this symbol you are to record all information that you need from your conversational partner.

Procedure

Read the information that you are given once before starting the conversation.

If you are you the caller, call your partner. Play through the scenario and then hang up the conversation. Now click "Next". First, you evaluate the overall quality of the conversation and then the conversation regarding the seven characteristics.

If you are you the one who gets called, please wait until you hear the ringtone. When you do, pick up. Play through the scenario with your partner and wait until they hang up the conversation. Click on "Next". First, you evaluate the overall quality of the conversation and then the conversation regarding the seven characteristics.

Phase 2 – Listening and speaking

Explanation

In this phase either you read two sentences to your partner or your partner reads two sentences to you. Afterwards you swap.

Procedure

You should have two sentences before you: call your partner and read both sentences. Hang up the conversation. Click "Next". Now, evaluate the overall quality of your speaking experience. Afterwards, evaluate your speaking regarding the two characteristics: "impact of your own voice" and "degradation of your own voice".

You should you have no sentences to read; wait until you are called and pick up the conversation. Listen to the two sentences and wait until your partner hangs up. Click "Next". Evaluate the "overall quality" of your listening experience, followed by the four conditions "noisiness", "coloration", "continuity" and "volume".

Phase 3 – Interaction

Explanation

In this phase, you and your partner are to perform a number verification.

Procedure

You will see four series of numbers before you. Two are **bold** and two are not. You are to read the bold number sequences to your partner. Should the first of the four sequences on your screen be bold, please call your partner. Perform the number verification in an alternating fashion and when all four sequences have been read, hang up the conversation. Wait after every number in the sequence (e.g., 24) for a confirmation from your partner in the form of a "yes" when they have the same number or a "no" if they have a different number. After the conversation, click "Next". Please evaluate the "overall quality" of the interaction and afterwards evaluate the characteristic "interaction".

Overall procedure

Summarize how the test goes, for every condition, in the following way (11 conditions in total):

call → phase 1 → hang up → Next → overall quality → evaluate 7 characteristics → call → phase 2 (speaking/listening) → hang up → Next → overall quality → 2/4 characteristics → call → phase 2 (speaking /listening) → hang up → Next → overall quality → 2/4 characteristics → call → phase 3 → hang up → Next → overall quality→ interaction.

With a few of the conditions you could have the feeling that you have evaluated them before. This is not the case. Please evaluate every condition independently from all your other evaluations. Try not to remember how you evaluated other "similar" conditions before; instead evaluate each characteristic of every condition individually.

Please evaluate the scales quickly and intuitively. This experiment is purely subjective in its nature. There are no right or wrong answers. **The system, not you, is being tested**. Only your personal opinion matters for this experiment.

If you have any additional questions, please don't hesitate to ask the experiment supervisor. Have fun! ☺

Appendix I

Examples for the speaking part of the test methodology – translated from German

(This appendix does not form an integral part of this Recommendation.)

"Can you please give me the best connection between Munich and Berlin? I have to arrive on Saturday at 12.30 pm latest."

"Please connect me with the complaints department. The repair on the main drain of my house has not been carried out properly."

"Here is the local fire station. We are trying to find an emergency call that has been suspended without giving any specific information. It sounded like a local call."

"I'm really sorry that I could not come to dinner at the weekend. Just before I wanted to leave, I had a small accident."

"Dear mother, here in Palma it is beautiful. The weather is hot and sunny and the sea is unbelievable. This morning I made a hike."

"On warm nights, I often lie in my bed and look through the open window. The moon shines down on me and it looks as if he smiles."

"Tomorrow morning my daughter wants to bake a pie. She insists on making everything by hand. She is convinced that it tastes much better then."

"What am I going to eat tonight? I still have a stew in the freezer, but the problem is that I eat it two to three times a week."

"Please connect me with the repair department. My TV has been with them for three weeks now and I want to know when it's finally done."

"I hate it when it rains on Monday. The streets are slippery and I have to be very careful when I walk to the station."

Appendix II

Results from an initial pilot test and second retest

(This appendix does not form an integral part of this Recommendation.)

To evaluate the test methodology a pilot test was conducted as described in [b-Köster2017]. Here, eleven test conditions were tested and analysed. The results showed a high reliability but also a need for analysing the inter-rater agreement of a second retest. For the latter, a retest was conducted [b-Köster2017]. Both tests provide subjective scores from the same lab, in the same country, using the same language and the same test conditions. In both tests, 36 naïve test subjects participated.

Figures II.1 and II.2 show examples of phase quality scores for the elven test-conditions derived from the presented test methodology in a similar lab in Germany using German (the data is taken from [b-Köster2017]). The two figures illustrate that subjects in two independent tests show a high degree of agreement on the phase quality scores for a common set of eleven conditions. For illustrative purposes, Figures II.1 and II.2 only show scores for the phase quality scores, but there was very good agreement for the dimension quality score as well. Table II.1 shows the correlation between the pilot and the retest scores for the sets of eleven conditions.

Typical results for the test methodology should include means and standard deviations for each condition in the test computed over the number of votes per condition.

Additional statistical analyses could include comparison of pairs of test conditions by student's t-test or evaluation of experimental factors by analysis of variance. Procedures and guidelines for the evaluation of individual subject's performance are included in the ITU-T Handbook "Practical procedures for subjective testing" [b-ITU-T_Testing].

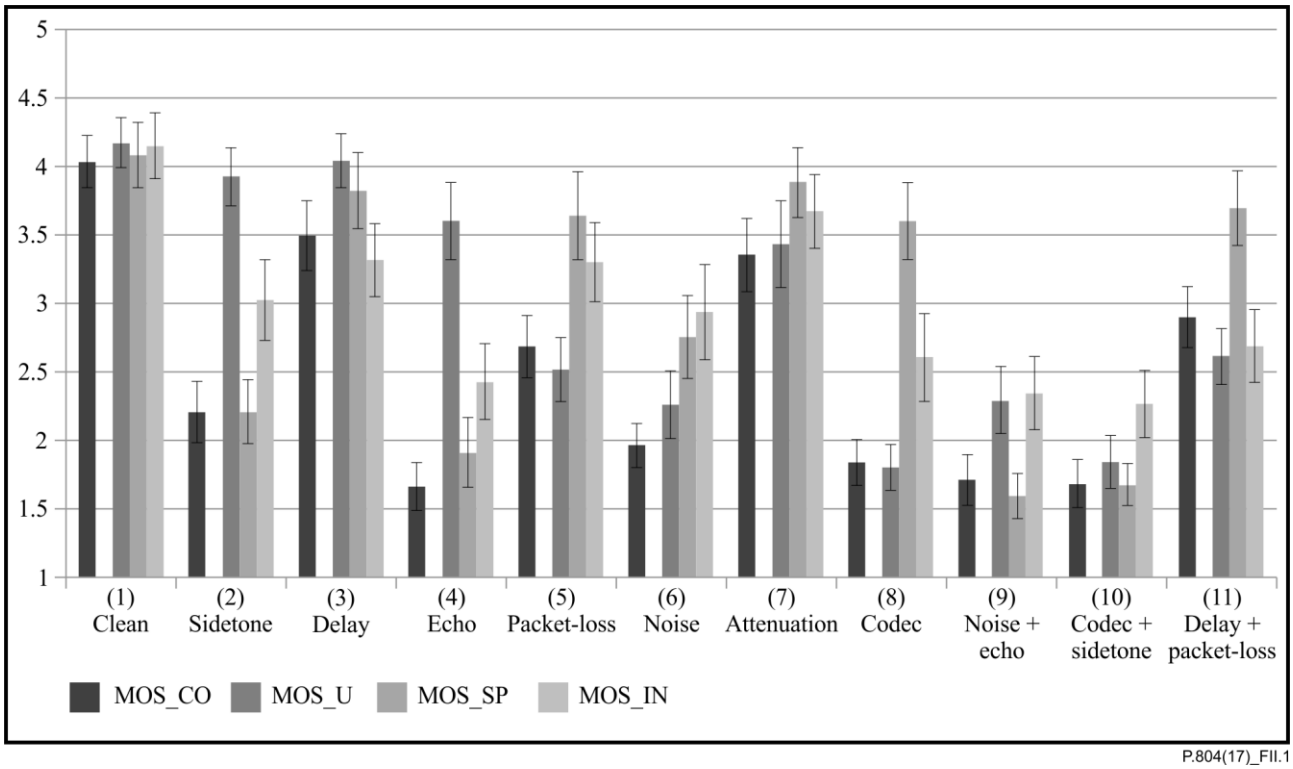


Figure II.1 – Subjective quality ratings resulting from the first pilot test; the overall conversational quality (MOS_CO), and the quality of the three conversational phases (MOS_LI, MOS_SP and MOS_IN). The error-bars display the 95% confidence intervals

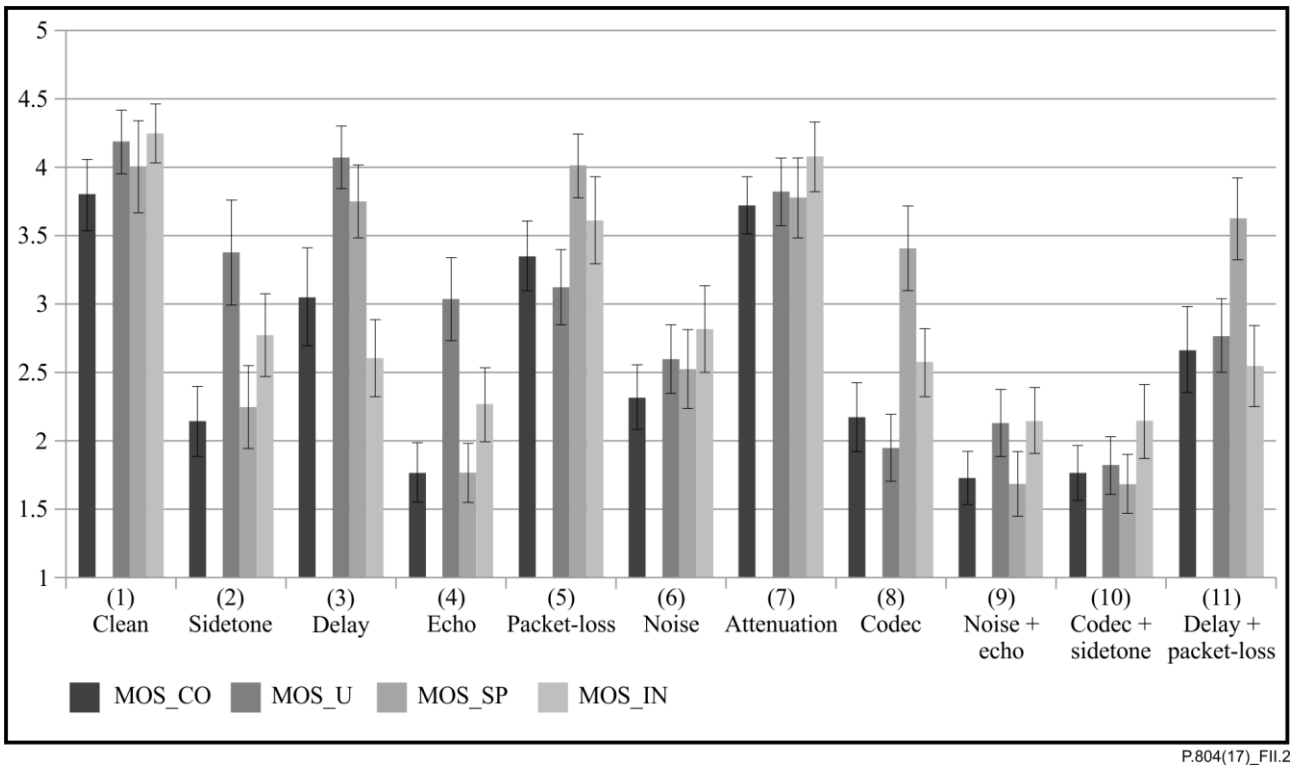


Figure II.2 – Subjective quality ratings resulting from the retest; the overall conversational quality (MOS_CO), and the quality of the three conversational phases (MOS_LI, MOS_SP and MOS_IN). The error-bars display the 95% confidence intervals

Table II.1 – Correlation between the pilot test and retest scores for the 11 conditions; CO – overall quality, LI – listening phase, SP – speaking phase, IN – interaction phase, Noi – noisiness, Dis – discontinuity, Col – coloration, Lou – loudness, Ios – impact of one's own voice on speaking, Dos – degradation of one's own voice, and Int – interactivity

MOS_CO	MOS_LI	MOS_SP	MOS_IN	MOS_noi	MOS_dis	MOS_col	MOS_lou	MOS_ios	MOS_dos	MOS_int
0.92	0.91	0.98	0.92	0.83	0.89	0.90	0.93	0.99	0.97	0.92

Bibliography

[b-ITU-T_Testing] ITU-T Handbook (2011), *Practical procedures for subjective testing*.

[b-Appel2002] R. Appel and J.G. Beerends, *On the Quality of Hearing One's Own Voice*, *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237-248, 2002.

[b-Guéguin2008] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, *On the Evaluation of the Conversational Speech Quality in Telecommunications*: EURASIP J.Adv. Signal Process, 2008.

[b-Köster2014] F. Köster and S. Möller, *Analyzing perceptual dimensions of conversational speech*, in Proc. *INTERSPEECH*, Singapore, Singapore, 2014, pp. 2041-2045.

[b-Köster2015a] F. Köster and S. Möller, *Perceptual Speech Quality Dimensions in a Conversational Situation*, in Proc. INTERSPEECH, Dresden, Germany, 2015, pp. 2544-2548.

[b-Köster2015b] F. Köster, D. Guse, M. Wältermann, and S. Möller, *Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech*, in *Fortschritte der Akustik – DAGA 2015: Plenarvortr. u. Fachbeitr. d. 40. Dtsch. Jahrestg. f. Akust*., Nürnberg, 2015, pp. 150-153.

[b-Köster2017] F. Köster, *Multidimensional Analysis of Conversational Telephone Speech*, Springer, Berlin, 2017.

[b-Wältermann2010] M. Wältermann, A. Raake, and S. Möller, *Quality Dimensions of Narrowband and Wideband Speech Transmission*, Acta Acustica united with Acustica, 2010, pp. 1090-1103.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
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	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
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, and associated measurements and tests
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, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2017