International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.1310

(03/2017)

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

Telemeeting assessment

Spatial audio meetings quality evaluation

Recommendation ITU-T P.1310

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.1310

Spatial audio meetings quality evaluation

Summary

Recommendation ITU-T P.1310 concerns the quality assessment of telemeeting systems that apply spatial audio rendering techniques to facilitate communication between parties at remote locations.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.1310	2017-03-01	12	11.1002/1000/13181

Keywords

Performance measures, spatial audio, subjective testing, telemeeting.

* 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.1310

Spatial audio meetings quality evaluation

1 Scope

The focus of this Recommendation is on test methods involving test participants, which collect subjective ratings, task performance measures, and/or descriptors of communication aspects. For that purpose, this Recommendation extends the testing procedures for telemeeting systems described in [ITU-T P.1301] by specifying test methods dedicated to spatial audio telemeeting systems and – where appropriate – referring to stand-alone methods that are also suitable for spatial audio, which are [ITU-T P.1311], [ITU-T P.1312]. Further test methods related to spatial audio will be under study within ITU and this Recommendation will be updated with the relevant references when new methods are available.

In addition, this Recommendation also provides some guidance on testing spatial audio systems which include video communication. The focus is on the perceived alignment of the audio and the video source locations.

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.1301] Recommendation ITU-T P.1301 (2012), *Subjective quality evaluation of audio and audiovisual multiparty telemeetings.*

[ITU-T P.1311] Recommendation ITU-T P.1311 (2014), *Method for determining the intelligibility of multiple concurrent talkers.*

[ITU-T P.1312] Recommendation ITU-T P.1312 (2016), *Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance.*

[ITU-T P-Sup.26] ITU-T P-series Recommendations – Supplement 26 (2012), *Scenarios for the subjective quality evaluation of audio and audiovisual multiparty telemeetings*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

None.

5 Conventions

None.

6 General recommendations concerning assessment of spatial audio telemeeting systems

It is recommended that subjective quality evaluations of spatial audio telemeetings are carried out as much as possible according to existing test methods recommended by ITU-T and ITU-R.

It is also recommended that the purpose of the test is taken into account when selecting and, if necessary, adapting appropriate test methods. Test purposes differentiate, for example, by the following aspects:

- rendering scenarios (e.g., audiovisual or audio only systems);
- the focus on the system components under test (e.g., overall system or individual system components);
- the target variables (e.g., overall quality, timbral quality, spatial quality, cognitive load, or task performance);
- the desired measurement sensitivity (e.g., differentiating just spatial audio vs. non-spatial audio or differentiating several instances of spatial audio); or
- by the measurement paradigm (e.g., subjective ratings or performance measures).

In case of multiparty tests, it is recommended that [ITU-T P.1301] is consulted for multiparty-specific advice when selecting the appropriate test methods.

It is recommended that spatial-audio-specific considerations are taken into account when selecting and, if necessary, adapting appropriate test methods. Such considerations may be based on the spatial-audio-specific aspects described in this Recommendation.

The selection, and if necessary adaptation of an appropriate test method may be done using the guidelines provided in this present Recommendation.

Finally, practitioners should be aware that spatial audio telemeeting systems can and should be evaluated on several metrics. For that reason, it is recommended that practitioners evaluate which set of aspects is important for a complete evaluation of a system under test and choose the test methods accordingly.

7 Guidance on appropriate test method

It is recommended that the actual test purpose at hand is first analysed in terms of the decision criteria in clause 7.1. Then, the guidance in clauses 7.2 and 7.3 can be applied to identify the most appropriate test method.

7.1 Decision criteria

The choice for a particular test method described in this Recommendation depends on the following six decision criteria:

- Rendering scenario:

  To utilize their full potential, spatial audio rendering approaches differ between use cases, and thus the assessment methodology should be chosen accordingly. Two main rendering scenarios are considered :

  1 audio-only or audiovisual systems using spatial audio for improved communication experience, independently from the video signal;

  2 audiovisual systems showing multiple participants for which the rendered audio space is expected to be aligned with the visual location of the participants.

- Test Focus:

  Spatial audio telemeeting systems in real networks consist of more components than just the spatial audio capturing and rendering engines, e.g., codecs, bandwidth, noise and echo reduction. For that reason, the test method should be aligned with the test focus; that is the evaluation of:

  1 the overall system; or

  2 individual system components.

- Target variables:

  The media signal quality of spatial audio has two main aspects; its spatial quality component and its non-spatial quality component:

  – The spatial quality component refers to the quality features related to the spatial attributes of sound. Examples of such attributes are immersion, envelopment, localization blur and source width.

  – The non-spatial quality component refers to the quality features related to the non-spatial attributes of sound. Examples of such attributes are loudness, timbre, noisiness and sharpness.

  Furthermore, the added value of spatial audio in telemeetings goes beyond media signal quality since it benefits communicative aspects such as conversation flow, communication effort, cognitive load and task performance. As a result, the perceived overall quality may be different from the perceived media signal quality.

  For that reason, the selected test method should be appropriate for the desired target variable(s). The variables considered here are:

  1 Overall quality (consisting of media-signal quality and communication quality components).

  2 Media-signal-quality component (consisting of non-spatial quality and spatial quality components).

  3 Communication-quality component.

- Measurement sensitivity:

  Assessment methods that have been proposed and used in the state-of-the-art can be divided into two categories in terms of their sensitivity:

  1 Methods that are able to show the added value of spatial audio systems, i.e., non-spatial audio vs. spatial audio.

  2 Methods that are – in addition – able to distinguish different variants of spatial audio systems.

  This needs to be considered when selecting a test method.

- Test paradigm: listening-only tests, conversation tests

  Both, listening-only (non-interactive) tests and conversation tests can be used for evaluating spatial audio telemeeting systems.

Listening-only tests have the advantage of being simpler to implement and easier to conduct. They enable the assessment of many short stimuli and tend to allow test subjects to focus more on the quality assessment task at hand.

Conversation tests allow the experimenter to set the test subjects into a more natural and realistic situation by using a partial or a full communication system. These tests enable the evaluation of conversational aspects, beyond pure signal quality, and the assessment of all components and properties of a full system.

- Measurement paradigm:

  An additional selection criterion is the measurement paradigm of a test method, which needs to be aligned with the assessment goals.

  1 If the assessment goals are to evaluate the quality experience consciously expressed by subjects, then subjective ratings are appropriate, as subjects reflect on the quality during a test.

  2 If the assessment goals are to measure quality differences without having subjects consciously reflecting on quality, then other test paradigms are required. For the purpose of spatial audio evaluation, conversational analysis and performance measures are appropriate approaches.

7.2 Considerations on the test focus

7.2.1 General considerations

Full-working spatial audio meeting systems consist of a number of individual components that contribute to the system performance. Such performance factors include:

- level/loudness matching/management;
- reduction of unwanted environmental and nuisance noise;
- allowing all talkers to be heard including in double talk (full duplex);
- not suppressing quiet speech sounds during capture and/or mixing;
- high quality audio such as wideband or super-wideband;
- audibility and differentiation of multiple talkers, e.g., through spatial separation.

On one hand, a sufficient understanding of the performance contribution of those individual system components will enable developers to optimize individual components.

On the other hand, great care must be taken when investigating, and especially when decomposing, solution performance not to exaggerate one aspect of performance that cannot be delivered in isolation in a practical system.

For that reason, it is recommended that both test scenarios are considered: Investigate individual performance factors to understand their importance and testing the quality of the whole system.

7.2.2 Implications for testing whole spatial audio systems

When evaluating whole communication systems with test methods involving test participants, a number of aspects need to be considered that may influence the comparability and repeatability of results:

- dynamic/adaptive and non-linear behaviour of system components, even in laboratory settings;
- possible interaction effects of individual components.

Focusing on spatial audio meetings, the following additional aspects should be considered, though these are not fully understood yet:

- interaction of benefit of spatial audio representation vs. impairments caused by non-spatial audio components, e.g., packet loss, noise in microphone signals;
- aspect of headphone compensation;
- multiple speakers in one room captured with one microphone.

It is recommended that such influences on comparability and repeatability are controlled as much as possible, unless the test purpose specifically requires that the system be evaluated without such control, like for instance final evaluation campaigns before product/service release.

The control of such influences may include technical modifications of the system under test, specific experimental designs that enable sufficient repetitions for averaging out unwanted random effects, or logging of technical information that can be used for data analysis.

If a control of such influences is not possible, feasible or intended, then it is recommended that such influences are checked during the data analysis to properly document any such effects found and to consider them when interpreting the results.

7.2.3 Implications for testing individual system components

When it comes to testing individual components of a spatial audio system, it is important to distinguish between testing those system components that are concerned with the spatial representation and those that are concerned with the non-spatial aspects of a telecommunication system.

The first category refers to aspects such as spatial sound capture, signal enhancement using spatial information (e.g., beam-forming) or addressing spatial aspects (e.g., multipath echo cancellation), spatial audio encoding, and spatial sound reproduction. The second category refers to aspects such as non-spatial signal enhancement (e.g., noise reduction) and signal transmission (e.g., delay).

For the testing of system components of the first category the methods of this Recommendation should be used. For the testing of system components of the second category, however, conventional non-spatial methods such as [ITU-T P.800] may also be considered if they have advantages in terms of test effort, sensitivity or other such aspects. Such considerations are valid as long as no mixed spatial and non-spatial conditions are used in the test (see, e.g., [b-Skowronek2015a] for a discussion on this), and as long as no interaction of the spatial system components and the non-spatial system components under test are expected. However, since practical experience is limited in the second respect, it is recommended that a pilot test is run to verify whether such interaction effects take place.

7.3 Selection of individual test methods

There are five different methods considered in this Recommendation that are presented in Annexes A to E, which address the different selection criteria described in clause 7.1. Table 1 provides an overview of the method that covers each selection criterion. It is recommended that this Table is used for selecting an appropriate method for the test purpose at hand. For that purpose, this table may be used as a checklist: the chosen method should have a checkmark (✔) in this table for each considered selection criterion.

Table 1 – Overview of how the five different methods of Annexes A to E cover of the different selection criteria of clause 7.1

Method			Annex A	Annex B	Annex C	Annex D	Annex E
Rendering Scenario	Audio-only or audio component of an audiovisual system		✓	✓	✓	✓	✓
	Audiovisual		✓	✓			
Test focus	Overall system		✓	✓	✓	✓	✓
	Individual components		✓	✓	✓	✓	
Target variables	Overall quality		✓	✓		✓ (3	✓ (3
	Media quality	Non-spatial quality	✓				
		Spatial quality	✓	✓			
	Communication quality		✓	✓	✓ (2	✓ (2	✓ (2
Measurement sensitivity	Sufficient (test added value of spatial audio)		✓	✓	✓	✓	✓
	High (differentiate variants of spatial audio)		✓ 1)		✓	✓	
Test paradigm	Listening-only (Non-interactive) Test		✓	✓	✓		✓
	Conversation test		✓	✓		✓	
Measurement paradigm	Subjective ratings		✓	✓			
	Performance measures				✓	✓	✓

NOTES:

1) Depending on the actual implementation of the method details, e.g., training phase, it is possible to achieve high measurement sensitivity.

2) These methods (indirectly) measure individual aspects of communication quality.

3) These methods are not measuring overall quality, but they may be combined with subjective ratings of overall quality.

Annex A

Method to collect subjective quality ratings for spatial audio meetings

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

This annex describes in more detail the set-up of a spatial-audio meeting assessment test that collects quality ratings from test participants, either by means of a non-interactive (listening-only) test, or conversation test.

It follows the general approach of [ITU-T P.800] and [ITU-T P.805] by addressing test facilities, conversation tasks and non-interactive stimuli, experiment design, test subjects, scales, instructions and training phases, data collection and analysis.

This annex also explains in detail spatial-audio specific aspects that need to be considered while it refers to [ITU-T P.800] and [ITU-T P.805] for details that are not spatial-audio specific, and to [ITU-T P.1301] for details that are multiparty-specific.

A.1 Test facilities

It is recommended that test facilities are used according to [ITU-T P.800] for two-party and [ITU-T P.1301] for multiparty scenarios, unless

– this annex is combined with another method of evaluating spatial audio meetings and that other method requires other characteristics, or

– aspects of the test facilities (e.g., room noise or reverberation) are test factors, or

– the test scenario specifically aims at realistic communication environments which deviate from the recommended test facilities.

If the sound reproduction devices are not part of the system under test but part of the test facilities, the product name of the devices or ideally their characteristics, in particular frequency response, should be documented.

A.2 Conversation tasks and non-interactive stimuli

A.2.1 Conversation task

It is recommended that conversation tasks are used according to [ITU-T P.805] for two-party and [ITU-T P.1301] for multiparty scenarios, unless

– this annex is combined with another method of evaluating spatial audio meetings and that other method requires other conversation tasks; or

– other conversation tasks are test factors; or

– the test scenario requires specific tasks that are not included in [ITU-T P.805], [ITU-T P.1301] or other annexes of this Recommendation.

When choosing one particular conversation task, the following characteristics should be considered:

– Multi-talk: How much multi-talk (cross-talk, talker overlap) is triggered by the conversation task?

– Conversational structure: How structured is the conversation, i.e., how predictable is the order when speakers contribute?

– Interactivity: How interactive, in terms of amount and speed of speaker changes is happening?

– Content relevance: Does the task require participants to do something with the content that the interlocutors say, e.g., writing minutes, answering questions after the call, etc.?

– Required level of familiarity (see also clause A.3 test participants): Does the task require groups of test participants that are familiar with each other?

Those aspects are particularly relevant for spatial audio meeting scenarios since they may influence the test participants' ability or need to separate and follow individual speakers, and thus the benefits of the spatial audio meeting system under test.

A.2.2 Non-interactive stimuli

In terms of the content, it is recommended that stimuli according to [ITU-T P.800] or [ITU-T P.1301] are used, unless:

– this annex is combined with another method of evaluating spatial audio meetings and that other method requires other stimuli; or

– different types of stimuli are test factors; or

– the test scenario requires specific stimuli that are not included in [ITU-T P.800], [ITU-T P.1301] or other annexes of this Recommendation.

When choosing a particular type of stimuli, the following characteristics should be considered:

– The number of communication aspects: Are the stimuli excerpts from conversations (according to [ITU-T P.1301]) or unrelated sentences (according to [ITU-T P.800])?

  Using excerpts from conversations has the advantage of better reflecting communication aspects in the participants' ratings, but the disadvantage that stimuli are required to be longer than a few seconds. Here good experience has been made with stimuli that allow about 20 seconds [b-Skowronek2015c] to one minute per interlocutor [b-Skowronek2015a]. Using unrelated sentences has the advantage of allowing for more stimuli per test session, but do not require test participants to take communication aspects into account. On the one hand this may let test participants focus on technical/signal aspects only but on the other hand those communication aspects may be necessary to properly reflect the added value of the spatial audio system under test.

– Mix of voice characters (gender mix) of speakers in the stimuli:

  Do the speakers in the stimuli have spectrally separated voices (as it is often the case in mixed-gender groups) or spectrally similar voices (as it is often the case in unisex groups) that could influence the test participants' ability (or need) to separate and follow individual speakers? This in turn influences the benefit of the spatial audio system: the benefit is expected to be stronger for similar voices (same genders) than for different voices (mixed genders).

– Placement of speakers in the virtual acoustic space:

  If this placement is not fixed or pre-defined by the system under test, the positioning of the speakers should be well defined, since the benefit of spatial audio is influenced by this positioning. One aspect is the distance between speakers, which can in principle range from zero (which is achieved by positioning all speakers on the same position) to a maximum possible distance (which is achieved by distributing all speakers with maximal distances between them). Another aspect is the symmetry or asymmetry of the positioning around the centre front direction, which may also influence the participant judgments. Related to this is the question of whether the dominant speakers are placed (more) to the central position or not.

– Temporal relation of speakers in the recordings:

  The amount of speech overlap between talkers is an important aspect for testing a system's capability to support multi-talk. Furthermore, the patterns of changes between single-talk and multi-talk states may also influence results.

– Suitability of stimuli for subtle differences between system configurations:

  In test scenarios investigating subtle differences between system configurations, the stimuli should allow the test participant to perceive such subtle differences. In particular the combination of the above mentioned aspects should be validated in that respect, for instance by means of a pilot test.

A.3 Test participants

It is recommended that special care is taken in the selection of test participants, taking into account the test scenario at hand. When deciding on a (set of) subject profile(s), and in case of conversation tests when pairing them into the test groups, the following aspects should be considered:

– Level of experience of the participants with specific equipment or technology: Following the advice in [ITU-T P.805] and [ITU-T P.1301], the participants' experience with spatial audio technology is of particular interest. At the time of writing this recommendation spatial audio meeting systems are still rather new to the general public, which could mean that fully naïve participants may be either over-optimistic due to a "wow" effect or over-pessimistic due to a "strangeness"-effect.

  The desired level of participant experience depends on the test purpose at hand; however, it is recommended that test participants with different levels of experience are not mixed. A preliminary training stage before the test can be used to achieve a more similar level of experience with the system/conditions under test (see below).

– Level of familiarity of the participants within a conversation group:

  The aspect of whether participants know each other or not could influence the test participants' ability (or need) to separate and follow individual speakers. This in turn could influence the added value of spatial audio rendering and is essentially influencing the measurement sensitivity (see clause 7.1).

  It is recommended that participants are chosen based on whether they are familiar with each other or not, unless:

  - this annex is combined with another method of evaluating spatial audio meetings and that other method requires mixes of familiarity; or
  - familiarity is a test factor.

  Another aspect is the known issue that in conversation tests with rather highly interactive tasks, it is hard for subjects to get into a fluent conversation if they do not know each other.

– Voice characters of test participants in conversation tests:

  The aspect whether conversation groups contain speakers with spectrally separated voices (as it is often the case in mixed-gender groups) or spectrally similar voices (as it is often the case in unisex groups) could influence the test participants' ability (or need) to separate and follow individual speakers.

  It is recommended that a balanced mix of groups is used with spectrally separated and spectrally similar voice characters (or mixed-gender and unisex groups), unless:

  - this annex is combined with another method of evaluating spatial audio meetings and that other method requires a different distribution of separated and similar voice characters; or
  - voice character is a test factor.

– Special populations, e.g., children, people with hearing or speech impairments:

If the test is targeted to special populations, additional characteristics may need to be considered, possibly even with such priority that (some of the) above mentioned advices cannot be fully considered.

A.4 Experiment design

It is recommended that typical experimental designs, such as balanced designs or randomization are used. The choice for a particular design type should take into account how far the above described influencing aspects (conversation test tasks, non-interactive stimuli and participant profiles) should be addressed in the design. For example, if there is no external reason to decide for one type of voice character (e.g., only male, or only female speakers), a design which balances the number of male-only and female-only groups should be considered.

A.5 Scales

It is recommended that quality judgments using the established rating scales of [ITU-T P.800] and [ITU-T P.805] are used.

Other questions may be added that specifically address individual aspects that are relevant for spatial audio meetings. In this regard, the investigator should strive for a balance between the number of individual aspects one could ask and keeping the test as simple as possible for participants, i.e., limiting the number of questions in a test. While a generally valid absolute number of questions cannot be given, a typical number of additional questions are in the order of three to ten for test protocols with naïve participants (e.g., in multidimensional analysis of quality: [ITU-T P.806], [b-Wältermann2013], [b-Köster2015], in studies on spatial audio quality: [b-Skowronek2015a], in studies on spatial attributes of room acoustics: [b-Berndtsson1994]).

There are two types of additional questions that may be considered: those that address individual communicative aspects that are relevant for spatial audio meetings and those that address individual signal and system characteristics.

Concerning questions on communicative aspects, examples are:

- Concentration effort: "It required (extremely much . . . extremely little) concentration to follow the conference."
- Speaker recognition effort: "During the conference, it was (extremely difficult . . . extremely easy) to recognize who was speaking."
- Topic comprehension effort: "It was (extremely difficult . . . extremely easy) to follow which opinions were exchanged during the conference."

Concerning questions on individual signal and system characteristics, there is currently no agreement on the exact number of dimensions and aspects that should be used: ITU-R Recommendations BS.1116 and BS.1534 provide rather rudimentary lists of attributes related to spatial audio; and multiple proposals for more comprehensive attribute lists are currently under study (e.g., [b-Lindau2014], [b-Zacharov2016]).

Current research on this topic suggests (e.g., [b-Rumsey2005], [b-Lindau2014], [b-Zacharov2016]), that spatial audio systems can be characterized by spatial and non-spatial attributes. Thus the investigator needs to decide whether non-spatial attributes should be included into the system assessment or not.

Until some agreed-upon set of items is available, it is recommended that state-of-the-art methods from multidimensional quality assessment (e.g., [b-Wältermann2013], [b-Köster2015]) are applied in order to find an adequate set of dimensions. Example methods are attribute elicitation, attribute scaling (using semantic differentials) followed by factor/principal component analysis, or pairwise comparison followed by multidimensional analysis. These methods have been particularly used for non-expert listeners. Alternative methods apply focus groups with experts (e.g., [b-Lindau2014]) or ratings obtained from trained listener panels (e.g., [b-Zacharov2016]).

A.6 Instructions to subjects and training phases

It is recommended that [ITU-T P.800], [ITU-T P.805] and [ITU-T P.1301] are consulted for advice on instructions and training phases, unless

– this annex is combined with another method of evaluating spatial audio meetings and that other method requires specific instructions and training procedures; or

– the mentioned advices are contradictory to the considerations discussed in the following.

Concerning instructions, the level of detailed explanations should reflect the level of detailed aspects that the investigator is aiming at. For instance, if test participants should judge only on the spatial quality or should judge both spatial and non-spatial quality, then the instructions should highlight that these different aspects exists. However, to avoid confusion such instructions should not become too complicated. Subjects should be asked after the training phase, if they understood the instructions as intended. This may be achieved by using specific stimuli in the training that clearly differ in the individual aspects and by asking the subjects or inspecting their ratings, how they judged those stimuli.

However, if a more holistic overall quality judgment is of interest, then such detailed explanations are not necessary.

Concerning the training phase, two effects are of particular interest: the effect of training on participants' experience with the equipment or technology under test (see subject profiles above) and the effect on measurement sensitivity.

With regard to experience, it has been observed that inexperienced listeners need a longer training phase in order to be able to perceive different spatial qualities. For that reason, it is recommended either to plan for such an extended training phase in case of test participants that have very little or no experience with spatial audio; or to limit the subject profile to participants that already have some experience.

With regard to measurement sensitivity, naïve test participants may be helped to focus on the relevant aspects by presenting example stimuli or by having training calls in example conditions. While this is a common procedure, one way of increasing measurement sensitivity is to further extend the training. One possibility that may be considered is to present at least one example for every test condition. For instance, it has been observed in a listening-only test that two specific examples for each condition during the training was sufficient to enable naïve participants to differentiate between different variants of a spatial audio system. However, until more practical experience on the required exact amount of training is available, it is recommended that a pilot test is conducted to find the proper number of stimuli for the training phase.

Annex B

Assessing the accuracy of the spatial alignment of audiovisual systems using spatial audio

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

B.1 Introduction

This annex concerns test scenarios in which the spatial alignment of the audio and video cues in the telemeeting system is of interest. It provides guidance on the measurement of the perceived localization of the conversation partners in auditory and visual space. There is no recommended test method for the assessment of the alignment of video and audio scenes at the time of writing. Until such method is developed, general advices are given for the assessment of audiovisual scene alignment. The importance of an accurate alignment and the implication of misalignments on communication effectiveness, cognitive load, discomfort and overall experience are for further study.

B.2 Evaluation criteria concerning audiovisual alignment

The quality assessment of audiovisual alignment should consider the following criteria:

1) Quality of spatial alignment of the visual and auditory scene (see Figure B.1)

   a) Evaluation of correctness of audio source localization, in terms of whether participants are placed at sufficiently correct angles.

      Example measures:

      i) Subjective ratings of alignment, such as "How correct is the alignment between the speaker positions that you hear and that you see?" – (1: very incorrect / 2: rather incorrect / 3: neutral / 4: rather correct / 5: very correct).

      ii) Localization performance, e.g., in terms of perceived number of matches and mismatches of the precise auditory and visual positions, or in terms of perceived distance between auditory and visual positions (e.g., b-DeBruijn2004).

   b) Evaluation of correctness of spatial ordering, in terms of whether participants are correctly placed from left to right, but without aiming for correct angles. Example measures:

      i) Subjective ratings of alignment, such as "How adequate is the spatial ordering of the speaker positions that you hear?" – (1: very inadequate / 2: rather inadequate / 3: neutral / 4: rather adequate / 5: very adequate).

      ii) Localization performance, e.g., in terms of perceived number of matches and mismatches of the spatial ordering (from left to right).

2) Quality of temporal alignment of visual and auditory scene

   Aspects include: timely video switching of active speakers, lip sync.

3) Interaction with non-alignment related criteria, such as:

   a) Quality of visual scene rendering

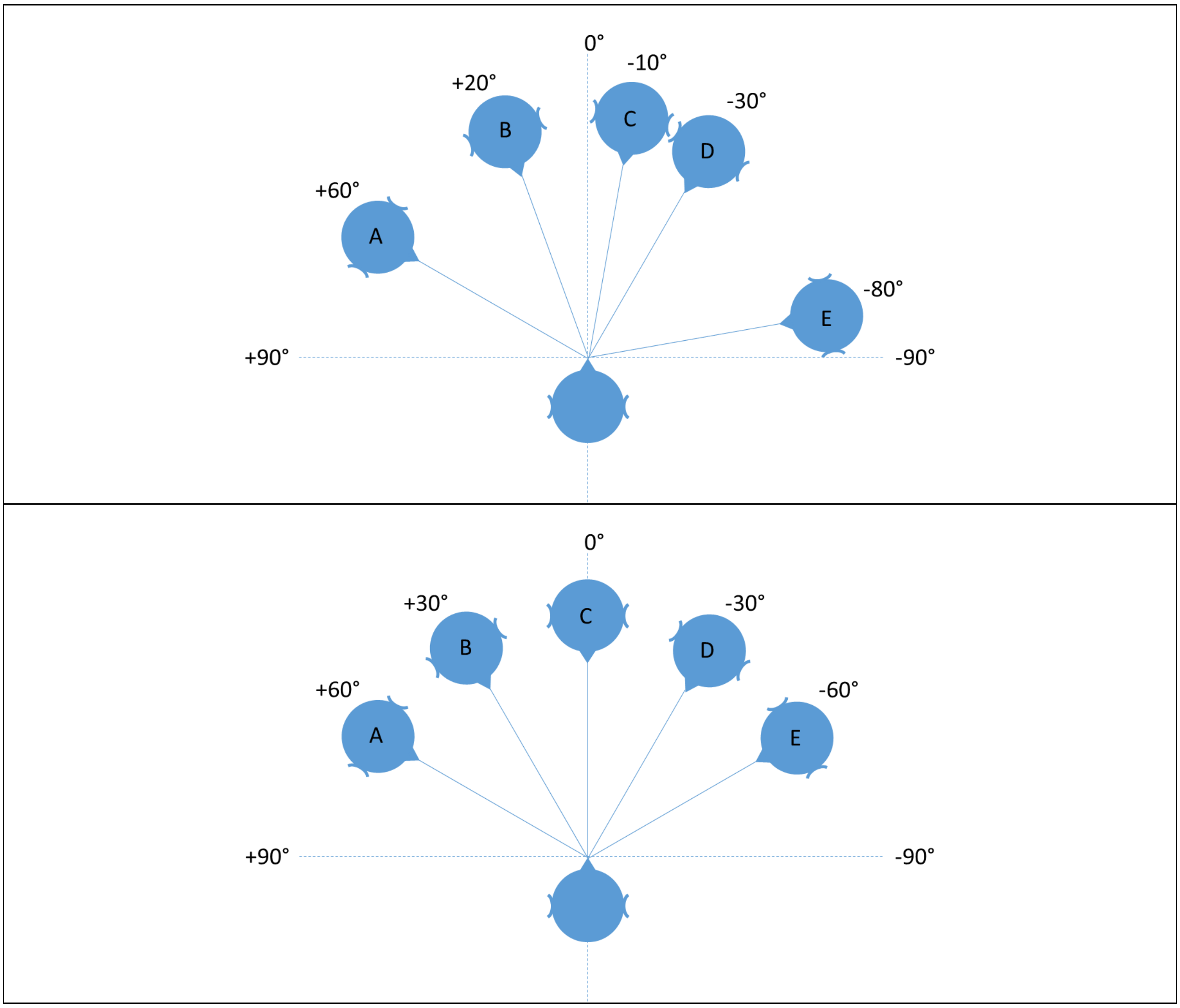
      Aspects include: Video quality of individual videos, visual depth (3D video), video layout, eye contact, visual scene dynamics, video switching.

   b) Quality of auditory scene rendering

      Aspects include: Non-spatial quality of individual audio signals, and spatial quality aspects.

c) Ease of communication

Was it easy to follow who was saying what? Was it possible with little cognitive effort? Was the interactive responsiveness good?



Top panel: Positions of speakers are placed in the auditory scene at those angles that are given by a visual scene. Bottom panel: The speakers are placed at other angles, but maintain the spatial ordering in terms of "from left to right".

Figure B.1 – Different options for aligning visual and auditory scene

B.3 Rendering scenarios affected by audiovisual alignment

It is recommended that an evaluation test focuses on one of the following three telemeeting rendering scenarios, since the evaluation criteria slightly differ between these scenarios.

B.3.1 Video conference with full spatial alignment between visual and audio scenes

Example use cases

A broad visual scene in terms of the angle span as observed from the user(s).

Example systems

1 A telepresence room with three adjacent screens and the users sitting 3to 4 meters from the screens with spatial audio rendering through a loudspeaker configuration of one or more loudspeakers.

2 A single user sitting in front of a wide computer screen at a distance of approximately 0.5 meters with spatial audio rendering through a loudspeaker configuration of one or more loudspeakers.

3 A single user sitting in front of a wide screen with spatial audio rendering through headphones.

Recommended set of evaluation criteria according to clause B.2

The users would expect an exact spatial alignment between the visual and auditory scene.

Thus, the following criteria should be considered: 1a, 2, 3a, 3b, 3c.

B.3.2 Video conference with partial spatial alignment between visual and audio scenes

Example use cases

A narrow visual scene in terms of the angle span as observed from the user(s), in which the screen shows multiple remote locations simultaneously, e.g., from left to right.

Example systems

1 A conference room with a single screen and the users sitting 2 to 5 meters from the screens with spatial audio rendering through a loudspeaker configuration of one or more loudspeakers.

2 A single user sitting in front of a smart phone or a tablet mounted in a dock with spatial audio rendering through a loudspeaker configuration of one or more loudspeakers.

3 A single user sitting in front of a smart phone or a tablet with spatial audio rendering through headphones.

Recommended set of evaluation criteria according to clause B.2

Even though the users may not expect an exact spatial alignment between the visual and auditory scene, they might expect a plausible spatial ordering of participants in the horizontal plane.

Thus, the following criteria should be considered: 1b, 2, 3a, 3b, 3c.

B.3.3 Video conference with no spatial alignment between visual and audio scenes

Example use cases

A narrow visual scene in terms of the angle span as observed from the user(s), in which the visual scene on the screen is rather simple and a decision was made not to spatially align the audio space with the visual space.

Example systems

1 A conference room with a single screen and the users sitting 2 to5 meters from the screens with spatial audio rendering through a loudspeaker configuration of one or more loudspeakers.

2 A single user sitting in front a smart phone or a tablet mounted in a cradle with spatial audio rendering through a loudspeaker configuration of one or more loudspeakers.

3 A single user sitting in front of a smart phone or a tablet with spatial audio rendering through headphones.

Recommended set of evaluation criteria according to clause B.2

In this scenario there is no need to test the spatial alignment. Thus, the following criteria should be considered: 2, 3a, 3b, 3c.

Annex C

Listening test method to obtain task performance indicators for the intelligibility of concurrent speakers

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

[ITU-T P.1311] presents a method to obtain an objective measure of how well a telemeeting system allows users to follow a conversation when talk spurts of several talkers coincide. The method comprises a listening-only test that involves test participants listening to several concurrent talkers, identifying one of them, and reporting what that talker said.

This method essentially provides a task performance measure targeting communication aspects.

Furthermore, this method has shown a high measurement sensitivity: It is, for example, sensitive to changes in the perceived angular separation of talkers and can be used to differentiate between alternative implementations of system components such as sound field capturing microphones or virtual spatial auditory displays.

It is recommended that the method described in [ITU-T P.1311] is used for investigating to which degree a spatial audio telemeeting system preserves the independence of voices and limits perceptual interference between voices.

This method is applicable only to the evaluation of two or more concurrent talkers; it is not applicable to the assessment of intelligibility under conditions where only one talker is active at any one time.

Annex D

Conversation test method for the measurement of communication effectiveness using task performance

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

[ITU-T P.1312] describes a test method for quantifying the effectiveness of telemeeting systems in conveying information in multiparty conversation scenarios. This method utilizes a predefined set of tasks designed to provoke rapid turn-taking and concurrent talking among participants. The goal of these tasks is to stress conferencing systems in order to clearly and measurably demonstrate their limits. The method measures the rate at which multiple participants exchange information to assess the effectiveness of communication systems compared to face-to-face communication.

Thus, this method essentially provides a task performance measure targeting at communication aspects.

Furthermore, this method has shown to provide stable and repeatable results across different laboratories and that is highly discriminative across various types of system properties, including duplex capability, spatialization, bandwidth and delay.

It is recommended that the method described in [ITU-T P.1312] is used for investigating to what degree a spatial audio telemeeting system is able to transmit multiple concurrent voices and maintain their independence.

This method applies only to voice communication; audio visual communications is not covered by the scope of this method.

Annex E

Listening test method to obtain task performance indicators for the cognitive load experienced in spatial audio meetings

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

E.1 Introduction and scope

Cognitive load refers to the limited capacity of the human working memory when performing a task. More specifically, cognitive load is constituted of three parts: intrinsic load, referring to the intrinsic nature of the task; extraneous load, referring to the instructions of the task; and germane load, referring to the capacity required for building up cognitive schemata for the task.

As an indicator for the cognitive load required to follow a telemeeting, this method measures the test participant's ability to memorize who contributed what information to the telemeeting by means of a memory test.

This method has been used in a number of studies investigating the added value of spatial audio rendering compared to non-spatial audio in conferencing scenarios [b-Baldis2001], [b-Raake2010], [b-Skowronek2015a]. The sensitivity of this method to distinguish different instantiations of spatial audio rendering is a topic for further study.

Furthermore, this method has been used for audio-only telemeetings; its applicability for other modalities such as audio-visual telemeetings or telemeetings providing graphical information means (e.g., screen sharing, web conferencing) is for further study.

E.2 Test method overview

The present method essentially applies a listening-only test paradigm: test participants listen to a number of speech stimuli and are asked to perform a memory test after each stimulus. During each memory test, participants are asked to judge for a number of transcribed quotations from the stimulus, which of the speakers made that statement.

E.3 Speech material

In contrast to conventional ITU-T P.800 tests, the speech material should not consist of short unrelated sentences. Instead, in line with [ITU-T P.1301], the stimuli should be recorded conversations. On the one hand, this enables the extraction of a sufficient number of quotations from the spoken content. On the other hand, the quotations per stimulus stem from the same conversation, which means that memory performance is measured for the same conversational context. In terms of recording length, good experience has been obtained with a minimum recording duration of about 1.5 to 2 minutes per speaker [b-Baldis2001], [b-Raake2010], [b-Skowronek2015a].

The recorded conversations may be free conversations if no strict control of the conversational complexity is needed. Alternatively, the recorded conversations may be structured conversations following defined scenarios, such as provided in [ITU-T P- Sup. 26] or [b-Skowronek2015b]. Using structured conversations is particularly recommended for studies, in which the effect of the number of participants is under investigation.

E.4 Quotations

One way to obtain suitable quotations is to first generate transcriptions of the speech material and then to select whole sentences from the transcriptions as quotations. The quotations should stem from passages in which no multi-talk is happening to ensure that the quotation is completely understood. Exceptions are back channels ("hm-hm", "yes", laughter, etc.) from other speakers as long as no crucial information of the target speaker is masked by the other speaker. In addition, the order in which the quotations are presented to the test participant should be the same as they were uttered in the recording.

Concerning the required number of quotations, a rough guidance can be drawn from [b-Skowronek2011] and [b-Skowronek2015a], which showed that 10 quotations per conversation rated by 13 participants in the test were not sufficient to yield significant differences between test conditions, while 16 quotations per conversation rated by 24 participants were able to yield significant differences between the same test conditions. If longer recordings are available, the number of quotations should be increased to further increase the robustness against measurement noise. For instance, [b-Baldis2001] and [b-Raake2010] used 26 quotations per conversation.

Test participants are not allowed to write down any notes during the calls. During each stimulus, however, test participants are provided with basic information about the speakers, such as name, role and affiliation.

E.5 Memory test questions

For each quotation per stimulus, it is recommended that the following question is asked:

> *"Which of the speakers made that statement?"*

and test participants are allowed to mark only one option

□ *Speaker Mr./Mrs. A*

□ *Speaker Mr./Mrs. B*

□ …

□ *"I don't know"*

where the number of answering options is determined by the number of speakers that participants can select from. Notice that this number does not necessarily need to be the number of speakers in the conversation, but that this number of options should be kept constant during the test.

The latter aspect is particularly important for tests with different numbers of speakers, for which it is recommended that the number of options is limited to the minimum number of speakers in the test. This enables to keep the *extraneous cognitive load* of the memory test constant (the "*instruction*" task of deciding between a number of options), while the *intrinsic load* of the memory test (the "*real*" task of memorizing which speaker contributed what) depends on the stimuli.

E.6 Test design

A within-subject design is the preferred design, in particular concerning the data analysis (see below); between-subject designs and mixed designs are nevertheless possible.

Since the present method is a memory test, each conversation may be presented only once per test participants.

A balanced distribution of the combinations of conversation recording and technical test condition is preferred; for this purpose, Latin-square or Greco-Latin square designs may be used.

E.7 Data validation

After the data has been collected from the test, it is recommended that the appropriateness of the selected quotations is first validated, and then the basic memory performance of individual test participants can be validated.

Concerning the validation of quotations, two types of quotations should be checked:

a) those for which the chosen speaker is correct for (almost) all test participants in (almost) all test conditions; and

b) those for which the chosen speaker is (almost) never correct for all test participants in (almost) all test conditions.

If such quotations are identified, the investigator should first check (again), whether those quotations have some special character (e.g., rather short or rather long quotation, pronunciation or limited intelligibility, characteristic expression used by one speaker) which differs from the other quotations. Then, based on this check, the investigator needs to decide, whether those quotations should be deleted from further analysis or not.

Concerning the validation of the test participants' memory performance, it is recommended that scores from test participants are checked to identify those with a substantial root mean square deviation from the general mean. Then, based on this check, the investigator needs to decide, whether the data of such participants should be deleted from further analysis or not.

All decisions for deleting data points and the corresponding reasons are to be reported in the test documentation.

E.8 Data analysis

Typical statistical analysis methods such as ANOVA with post hoc tests or planned comparisons, or – if necessary – corresponding non-parametric tests should be applied to identify significant differences between conditions.

Notice that in case of within-subject designs, the statistical analysis (e.g., repeated-measures ANOVA) is conducted "within test participants". This in turn does not require to normalize all results of each test participant to his or her basic memory performance.

However, if absolute performance values are of interest, the investigator should verify, if such normalization is appropriate, for instance when there is a large spread of basic memory performance across test participants.

Bibliography

[b-Baldis2001] Baldis, J.J. (2001). *Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences*. In: Proceedings of the ACM CHI 2001 Human Factors in Computing Systems Conference. Ed.: M. Beaudouin-Lafon and R.J. K. Jacob. Vol. 3(1), pp. 166–173.

[b-Berndtsson1994] Berndtsson, G., Krokstad, A. (1994), *A room acoustic experiment with an artificial reverberation system using wooden loudspeakers*. Acta acustica 2(1), 37-48.

[b-DeBruijn2004] De Bruijn, W. (2004), *Application of WaveFieldSynthesis in Video Conferencing*, PhD Thesis, Laboratory of Acoustical Imaging and Sound Control, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands.

[b-Köster2015] Köster, F., Möller, S. (2015), *Perceptual Speech Quality Dimensions in a Conversational Situation*, In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech2015). International Speech Communication Association. Dresden, Germany, Sep.

[b-Lindau2014] Lindau, A., Erbes, V., Lepa, S., Maempel, H. J., Brinkman, F., & Weinzierl, S. (2014). *A spatial audio quality inventory (SAQI)*. Acta Acustica united with Acustica, 100(5), 984-994.

[b-Raake2010] Raake, A., Schlegel, C., Hoeldtke, K., Geier, M., and Ahrens, J. (2010), *Listening and conversational quality of spatial audio conferencing*. In: Proceedings of the AES 40TH International Conference. Tokyo, Oct.

[b-Rumsey2005] Rumsey, F., Zielinski, S., Kassier, R., Bech, S. (2005), *On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality*, Journal of the Acoustical Society of America, Volume 118, Issue 2, DOI:10.1121/1.1945368.

[b-Skowronek2011] Skowronek, J. Raake, A. (2011), *Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing*. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech2011). International Speech Communication Association. Florence, Italy, Aug., pp. 829-832.

[b-Skowronek2014] Skowronek, J., Raake, A., Da Silva, A., Roegiers, D. (2014), *Quality Assessment of Spatial Audio Conferencing Systems from an End User Perspective*, in Fortschritte der Akustik – DAGA 2014, German Acoustical Society.

[b-Skowronek2015a] Skowronek, J., Raake, A. (2015), *Assessment of Cognitive Load,Speech Communication Quality and Quality of Experience for Spatial and Non-Spatial Audio Conferencing Calls*, in Speech Communication (66), pp. 154-175.

[b-Skowronek2015b] Skowronek, J. (2015), *xCT – Scalable Multiparty Conversation Test scenarios for conferencing assessment (Version 01)*. doi: 10.5281/zenodo.16138

[b-Skowronek2015c] Skowronek, J., Weigel, A., Raake, A. (2015), *Quality of Multiparty Telephone Conferences from the Perspective of a Passive Listener*, in Fortschritte der Akustik – DAGA 2015, German Acoustical Society.

[b-Wältermann2013] Wältermann, M. (2013), *Dimension-based Quality Modelling of Transmitted Speech*, Springer.

[b-Zacharov2016] Zacharov, N., Pike, C., Melchior, F., Worch, T. (2016), *Next generation audio system assessment using the multiple stimulus ideal profile method* In: Proceedings of 8th International Conference on Quality of Multimedia Experience (QoMEX 2016), Lisbon, Portugal.

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