INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.920

(08/96)

SERIES P: TELEPHONE TRANSMISSION QUALITY

Audiovisual quality in multimedia services

Interactive test methods for audiovisual communications

ITU-T Recommendation P.920

(Previously CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY

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

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

ITU-T RECOMMENDATION P.920

INTERACTIVE TEST METHODS FOR AUDIOVISUAL COMMUNICATIONS

Summary

This Recommendation is intended to define interactive evaluation methods for quantifying the impact of coding artifacts and transmission delay on point-to-point or multipoint audiovisual communications. This methodology is based upon conversation opinion tests, and can be considered to be an extension of the methods defined in, Annex A/P.800.

The following points will be taken into account in the subsequent clauses:

– tasks to be proposed to the assessors;

– methods and experimental design;

– questionnaires.

Source

ITU-T Recommendation P.920 was prepared by ITU-T Study Group 12 (1993-1996) and was approved under the WTSC Resolution No. 1 procedure on the 30th of August 1996.

FOREWORD

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the ITU. The ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Conference (WTSC), which meets every four years, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 1 (Helsinki, March 1-12, 1993).

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.

CONTENTS

Recommendation P.920

INTERACTIVE TEST METHODS FOR AUDIOVISUAL COMMUNICATIONS

(Geneva, 1996)

1 Introduction

These audiovisual interactive test methods are intended to quantify the impact of factors, such as coding artifacts and transmission delay, that may affect the ability to conduct an interactive audiovisual communication.

The efficacy of these tests strongly depends on the ability to reproduce in laboratory environments the conditions that are very close to the real situations. In this regard, particular care must be taken in choosing the tasks proposed to the subjects. In general, those tasks used in conversation tests for telephony assessment are not suited for audiovisual assessment because they often distract the subject's attention from the video screen. Therefore new tasks have to be developed following the criteria illustrated in this Recommendation.

Substantial work has been done in this area, although all aspects of audiovisual quality are not yet completely understood.

This Recommendation reflects the current early status of research on interactive audiovisual testing.

As progress on this work continues, understanding of these interactive test methods will no doubt improve. As new knowledge is attained, this Recommendation will be revised.

2 Scope

This Recommendation is intended to define interactive evaluation methods for quantifying the impact of coding artifacts and transmission delay on point-to-point or multipoint audiovisual communications. This methodology is based upon conversation opinion tests, and can be considered to be an extension of methods defined in Annex A/P.800.

This Recommendation does not cover topics that are already included in other Recommendations such as the objective measurements of the quality of a link, that are already defined in other Recommendations of the P-Series.

3 References

The following 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; all users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

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

– ITU-R Recommendation BT.812 (1994), *Subjective assessment of the quality of alphanumeric and graphic pictures in Teletext and similar services.*

4 Experimental design

4.1 Basic approach and factors to be investigated

In order to quantify the impact of factors, such as transmission delay, that may affect the ability to conduct an interactive communication, the approach proposed in this Recommendation is based on an active talker conversation assessment. Further, since it is necessary to express these opinions using a rating system, several single-stimulus rating scales are proposed.

4.2 Stimuli used in conversational tests

In general, in conversational opinion tests it is desired to minimize the artificiality of the environment. However, at the same time, it is necessary to invoke some method to stimulate interactive communication utilizing the conditions which are being evaluated. In telephony assessments, it is common to use a set of photographs, or some other form of printed material, to achieve this objective. In audiovisual terminal performance assessments, however, such mechanisms are likely to distract a participant's attention from the video screen, thus possibly leading to an unnatural mode of communication for this type of terminal.

For general applications, the following guidelines are provided for designing task-based tests [3]:

– the task should be designed such that, during their conversation, the subjects primarily maintain their attention on the audiovisual terminal;

– the task must have sufficient face value, that is, it must resemble real-life audiovisual communication to a sufficient degree. In particular, it is preferable that the task be performed by two subjects and not by one subject and an experimental leader;

– the task must yield reproducible quantitative results that represent adequate measures of communication efficiency. When time delays are involved, time measures should be among the results.

A wide range of subjects, including elderly and hearing-impaired subjects, should be able to perform the task.

It is preferable that the task is, in itself, sufficiently rewarding for the subjects. This has several advantages: the subjects learn the task faster and they are less susceptible to fatigue and loss of motivation.

From past experiments, it has been found that lively audiovisual conversations can be stimulated if the participants in such a test know each other. Subsequently, the provision of written material can be used as a secondary, rather than primary, source of stimulation. Thus, unlike telephony, familiarity between pairs of conversing participants is highly desirable, if not essential.

It is recognized, however, that for specific applications, the conversational tasks may have to be modified to take into account the services that the system under test is intended to provide. In order to permit meaningful measurements to be made of the factors being investigated, it is recommended that in such cases the conversational tasks be structured so as to represent the applications of interest, particularly as regards:

a) the rate of information exchange; and

b) the degree of audio and video signal utilization.

For example, to account for the attributes in the first category, tasks could range from predominately one-way communication, to free-conversation, to a rapid exchange of information, be it via video, audio or both signals. Similarly, to test attributes in the second category, tasks could range from the subjects working on a hard-copy document in front of them (minimum use of video information) to reading sign language over the video link (maximum use of video information). The actual tasks

should combine attributes from both categories. These guidelines have been applied to develop the tasks illustrated in Appendix I, and the protocols for the tasks are detailed in Appendix II.

4.3 Test conditions and experimental design

In general, at least one transmission impairment factor or test condition is likely to be evaluated in a test, in addition to a baseline (reference) condition where the impact of such factor is minimum (when using the reference condition, this should not be identified as such to the participants). However, because conversational tests are time-consuming, the total number of conditions ought to be reasonably constrained in order to minimize participant fatigue and maximize experimental accuracy. This requirement should be balanced against the need to ensure that the duration of each conversation/condition is at least five minutes long.

As with conversational tests (audio communications), a Latin or Graeco-Latin square may be found to be a suitable experimental design for this purpose [1]. In such case, the square's rows may be associated with the test participants and the square's columns with the order in which the conditions in the test are being presented.

Other treatments may also be appropriate depending on the factors being investigated. For example, past experiments have appeared to indicate that there may be an interaction between the audiovisual communication path quality and the perception of the impact of transmission delay. Consequently, it may be preferable to apply two treatments using a Graeco-Latin square design, so that the letters of the first alphabet are associated with different values of transmission delay and the letters of the second alphabet are associated with different image/voice coding rates.

Of course, other experimental designs including replicated block designs and Youden square designs may be suitable and could be left up to the experimenter to select in order to meet specific cost and accuracy objectives in view of the number of conditions of interest.

Also any possible effects related to the order in which the tasks are performed must be taken into account.

4.4 Subjects

At least 16 subjects should participate in a test, the exact number will be dictated by the experimental design and the accuracy required to the results. These subjects should be non-expert, and they should not be directly involved with either audio and/or video technology as part of their normal work.

Nevertheless, in the early phases of the development of audiovisual communications systems and in pilot experiments carried out before a larger test, small groups of experts (4 - 8) or other critical subjects can provide indicative results with sufficient reliability.

4.5 Subject training and reference connections

Before starting the experiment, a scenario of the intended application of the system under test should be given to the subjects. The range and type of impairments should be shown in a preliminary phase. During this phase, a first level of personal introduction may thus be allowed to take place over the communication link at the worst (or best) experimental condition, while further discussion pertinent to the tasks expected of the participants can be subsequently permitted at the best (or worst) experimental condition.

Again, as with the main test, the particulars of the conditions should not be revealed to the test participants.

4.6 Ambient room and equipment characteristics

The rooms' ambient noise characteristics should be representative of office environments, while the acoustic isolation between the two environments should be better than 60 dB. The sound characteristics of rooms employed in conversational assessments and found in Recommendation P.800 should also be applicable in this case.

The visual/luminance characteristics of the rooms, brightness of the video display, equipment configuration in relation to lighting source(s) and the participant(s) are, according to ITU-R Recommendation BT.812, as follows:

Viewing distance:	from 4H to 8H (Note)
Peak luminance:	from 70 cd/m^2 to 200 cd/m^2
Screen contrast ratio without background illumination:	from 30 to 50
Ratio of background luminance to maximum screen luminance:	~0.25
Illumination:	about 500 lux
General chromatically:	white

NOTE – H indicates the picture height.

The viewing distance should be defined taking into account not only the screen size, but also the type of screen, the type of application and the goal of the experiment. For a given screen height, it is likely that the viewing distance preferred by the subjects increases when visual quality is degraded. Concerning this point, the viewing distance should be predetermined for qualification tests. More critical distance should be used in optimization tests.

It is preferable to use the whole screen for displaying the scene. Nevertheless when, for some reason, the scene must be displayed on a window of the screen, the colour of the background in the screen should be 50% grey corresponding to Y=U=V=128 (U and V unsigned).

Other possible factors related to hands-free operation may be required to minimize uncontrolled experimental variations (for further study).

5 Solicitation of opinions

As with telephone conversational assessments, each participant should also be separately solicited for his, or her, opinion after the completion of each condition. It is preferable to structure in advance the type of questioning and minimize the number of questions posed after each condition and to minimize uncontrolled variations. A possible test condition questionnaire that could be used for this purpose can be found in Appendix III.

Several category judgement scales can be used to evaluate the audiovisual terminal performance. The sensitivity of these scales to various transmission impairment factors may be different.

Examples of scales that may be used for this purpose are:

- Overall Audiovisual Quality, Video Quality, and Audio Quality are generally assessed using the categories: Excellent, Good, Fair, Poor, or Bad.
- Effort Needed to Interrupt is generally assessed using the categories: No Effort, Minor Effort, Moderate Effort, Considerable Effort, or Extreme Effort.
- Communications Difficulty and Acceptability of Communication are generally assessed using a binary choice: Yes or No.

Although selection of any particular scale may be subject to the goals of an individual experiment, it is important that these scales (and their associated wording, or translation) be used consistently by different laboratories.

If the use of audiovisual terminals is novel for most participants, it is recommended that "Exit Questions" are presented after the last test condition has been rated. Such questions should attempt to capture any other factors that may have been inadvertently omitted from the experiment. A sample of such questions is given in Appendix IV.

Bibliography

[1] KIRK (R.E.): Experimental design – Procedures for the behavioural sciences, 2nd Edition, *Brooks/Cole Publishing Co.*, California, 1982.

[2] COM XII-85 – (July 1991), *Proposals of quality assessment method due to delay and acceptable delay time*, NTT.

[3] BRONKHORST (A.W.), VERHAVE (J.A.): The effect of audio-video desynchronization on communication efficiency in videotelephony, Study for PTT Telecom Netherlands, *TNO Institute for Perception*, Report IZF 1992 C-35, 1992.

Appendix I

Examples of tasks and stimuli for conversation

I.1 Stimuli for conversation

The following tasks differ from each other in the degree to which free conversation can occur [3]. The protocols for the tasks are described in more detail in Appendix II.

Task 1 The Name-Guessing task. The name-guessing task is a question-answer game performed according to a fixed protocol. This results in a very restricted conversation.

Task 2 The Story-Comparison task. In the story-comparison task, subjects have to discover a number of differences between two versions of a story. They are allowed unrestricted conversation. Prior to the test, both subjects have to read and memorize a short story.

Task 3 The Picture-Comparison task. In this task, the subjects have to memorize a picture and subsequently determine whether they were given identical or different pictures. Conversation is not restricted.

I.2 Tasks to evaluate the effects of speech delay on communication quality

In the following tasks the talk spurt increases from task 1 to task 6, whereas the conversation switching rate decreases [2].

1) take turns in counting;
2) take turns reading random numbers aloud as quickly as possible;
3) take turns verifying random numbers aloud as quickly as possible;
4) words with missing letters are completed with letters supplied by the other talker;
5) take turns verifying city names as quickly as possible;
6) determine the shape of a figure described verbally;
7) free conversation.

The previous tasks (with exception of task 1 and task 7) cannot be used for audiovisual quality evaluations because most of them require the subjects to concentrate their attention on a sheet of paper and not on the screen.

I.3 Tasks to evaluate the effects of audiovisual delay on communication quality

The following tasks should draw the attention of the assessors to the video signal:

1) one of the subjects shows and describes a plasting building block and the other one is required to reproduce it;

2) one of the two subjects shows some exercises for the alleviation of neck pain and the other one is required to reproduce them.

I.4 Task to evaluate the synchronization between audio and video signals

The following task is intended for drawing the attention of the assessors to the synchronization between audio and video signals:

– One person claps while the other checks the synchronization between the movement and the sound.

Task 1 is not suitable for assessing lip synchronization.

Additional tasks aimed at evaluating the synchronization between the two signals are for further studies.

Appendix II

Protocols for the stimuli for conversation

II.1 Protocol for the Name-Guessing task

In [3] the following protocol for the Name-Guessing task is used:

"The first task, the name-guessing task, is a question-answer game performed according to a fixed protocol. In this task, one of the subjects receives three pieces of information: first, either the word "Brand" or the word "Person", indicating whether it is a brand name or a name of a (well-known) person that should be guessed; second, in the case of a brand, a description of the product, and in the case of a person, his or her profession; third, the name to be guessed. Thus, the subject would, for example, be presented with the text:

– Brand;

– Cigarettes;

– Camel."

"In guessing the name of the brand or person, the second subject has to adhere to the following protocol: the first question should be "Is it a brand or a person?"; the second, in the case of a brand, "What is the product?", and, in the case of a person "What is the profession?"; subsequently, one guess can be made; when the guess is false, consecutive letters of the name should be asked for and a guess can each time be made; this continues until the name is guessed correctly or until the entire name has been spelled out. The subjects being interrogated are allowed to consult their text during the conversation. Most subjects do not find this necessary, however, because they had no trouble retaining the limited amount of information contained in it.

After the name is guessed, the total time required and the number of letters that are requested are scored. Several names, varying in length and difficulty, should be guessed in a given experimental

condition. A linear regression analysis can then be applied to the data so that estimates are obtained of the time required for a direct guess (zero letters suggested) and the extra time per letter."

II.2 Protocol for the Story-Comparison task

In [3] the following protocol for the Story-Comparison task is used:

"In the second task, subjects have to discover a number of differences between two versions of a story. They are allowed unrestricted conversation. Prior to the actual test, both subjects have to read and memorize a short story. They are given stories that are essentially the same but that contain a certain number of small but distinct differences. (For example stories of approximately 200 words, containing six differences.) In addition, they receive a list of questions about the story, which they have to answer for themselves. The questions are intended to improve retention of the story. After the memorization period, the subjects have to start a conversation with the aim of discovering as quickly as possible all differences between the two stories. They are not allowed to consult the text during the conversation. The subjects know how many differences there are and they get feedback informing them when they have detected a difference and how many differences they have detected thus far. The conversation continues until all differences are detected or until no differences have been detected for a certain time."

"The starting time of the conversation and the times at which detection of a difference occurs are registered. In the analysis, the time interval between the start and the detection of the first difference, and the intervals between subsequent detections are determined. These intervals are then ordered according to their duration. This is done because the time interval, i.e. the time required for the detection of the next difference, depends both on the strategy used by the subject pair and on where the differences are located in the text. Reordering the intervals reduces the variance introduced by these factors. In addition, it allows evaluation of the effect being investigated as a function of the duration of the interval."

II.3 Protocol for the Picture-Comparison task

In [3], the following protocol for the Picture-Comparison task is used:

"In this task, the subjects have to memorize a picture and subsequently determine whether they were given identical or different pictures. (Pictures of various subjects, such as landscapes, buildings and urban sites, can be used.) Conversation is not restricted. The subjects are not allowed to look at the picture during the conversation.

The results that are scored are the total time required by the subjects to reach a decision and whether or not the decision is correct. As subjects normally need more time when the pictures are identical than when they are different, the results for both cases have to be separated. Though this task is somewhat similar to the story-comparison task, there is a difference in the degree to which free conversation will occur. This is because, in comparing stories, the subjects tend to borrow phrases from the text, known to both of them, whereas the task of comparing pictures forces them to use their own wordings."

Appendix III

Test condition questionnaire

Questionnaire to be used after each test condition:

Administrator: I will ask you a few questions (again) relative to your opinion of the connection over which you just conversed. Are you ready?

1 How would you rate the overall audiovisual quality: (Circle one)

Excellent Good Fair Poor Bad

2 How would you rate the video quality of the connection: (Circle one)

Excellent Good Fair Poor Bad

3 How would you rate the audio quality of the connection: (Circle one)

Excellent Good Fair Poor Bad

4 How would you judge the effort needed to interrupt the other party (or parties)? (Circle one)

No effort Minor effort Moderate effort Considerable effort Extreme effort

5 Did you have any difficulty during the connection? (Circle one)

Yes No

6 Was the connection acceptable? (Circle one)

Yes No

Administrator: Thank you for your answers. We will now go ahead and re-establish our connection in a few minutes so that we can continue the experiment. As soon as the connection is set up, I shall let you know that you can go ahead and converse with the other party.

Appendix IV

Exit questions

Additional questions to be used at the end of the test:

Administrator: Since we are at the end of the test, I would like to ask you some supplementary questions. Are you ready?

Could you tell me, in order of importance, which of the following factors would you consider sufficiently important that you would like to see improved. (Evaluator will read a number of factors here.)

7a_____ 7b_____ 7c_____ 7d_____

Do you have any other comments relative to the entire test you wish to make known to us?

_____

_____

_____

_____

_____

Thank you for your cooperation in taking part in the test. I shall be over in a few minutes to wrap up the session and let you out of the building.

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Telephone network and ISDN
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media
Series H	Transmission of non-telephone signals
Series I	Integrated services digital network
Series J	Transmission of sound-programme and television signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound-programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminal equipment and protocols for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communication
Series Z	Programming languages