INTERNATIONAL TELECOMMUNICATION UNION

ITU-T P.832

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(05/2000)

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

Methods for objective and subjective assessment of quality

Subjective performance evaluation of hands-free terminals

ITU-T Recommendation P.832

(Formerly CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

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

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

ITU-T Recommendation P.832

Subjective performance evaluation of hands-free terminals

Summary

This ITU-T Recommendation describes methods and procedures for conducting subjective performance evaluations of hands-free terminals.

The use of hands-free terminals in communication has numerous advantages for the telephone users, especially for all "non-traditional" types of terminals such as car phones, computer/laptop-type terminals and others. Due to the complex acoustical situation a big variety of signal processing which may be non-linear and/or time variant is expected. ITU-T Recommendation P.340 describes measurement techniques for hands-free terminals, ITU-T Recommendation P.581 describes the use of the HATS for the evaluation of terminals, ITU-T Recommendations P.501 and P.502 describe measurement signals and analysis procedures. Using these methods a minimum performance of hands-free terminals should be ensured. However, there is always the possibility that those tests do not address fully the impact of all kinds of signal processing in a hands-free terminal and their impact on speech transmission quality.

Subjective testing is a commonly used method of assessing the performance of terminals, including digital speech codecs, voice-operated signal processing, echo cancellation, noise reduction and other types of signal processing. This ITU-T Recommendation defines methods for the subjective evaluation all kinds of hands-free terminals.

Source

ITU-T Recommendation P.832 was prepared by ITU-T Study Group 12 (1997-2000) and approved under the WTSC Resolution 1 procedure on 18 May 2000.

Keywords

Hands-free terminals, speech transmission quality, subjective performance.

FOREWORD

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

The World Telecommunication Standardization Conference (WTSC), 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 WTSC Resolution 1.

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

NOTE

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

INTELLECTUAL PROPERTY RIGHTS

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

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

CONTENTS

Page

ITU-T Recommendation P.832

Subjective performance evaluation of hands-free terminals

1 General

1.1 Scope

This ITU-T Recommendation describes procedures to be used to assess the subjective performance of hands-free terminals. The methods defined here may be used to assess the extent to which a hands-free terminal operates effectively for speech. This ITU-T Recommendation does not define specific values for hands-free terminal parameters (e.g. convergence time of echo cancellers) to yield satisfactory subjective performance.

The procedures defined here may also be appropriate for evaluating the subjective performance of other types of terminals and signal processing devices.

A complete subjective evaluation of hands-free telephones can be performed by the combination of three types of tests: conversational test, double talk test and third-party listening test (listening only test).

In general the evaluation of hands-free phones performance must take into account conversational interactions between subjects; a conversational test is the only type of subjective test which allows such an evaluation.

If a more detailed evaluation of a hands-free terminal is needed, it is recommended to perform double talk test and/or third-party listening tests additionally.

1.2 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.

[1] ITU-T Recommendation P.10 (1998), *Vocabulary of terms on telephone transmission quality and telephone sets*.

[2] ITU-T Recommendation P.340 (2000), *Transmission characteristics and speech quality parameters of hands-free terminals*.

[3] ITU-T Recommendation P.501 (2000), *Test signals for use in telephonometry*.

[4] ITU-T Recommendation P.502 (2000), *Objective test methods for speech communication systems using complex test signals*.

[5] ITU-T Recommendation P.51 (1996), *Artificial mouth*.

[6] ITU-T Recommendation P.56 (1993), *Objective measurement of active speech level*.

[7] ITU-T Recommendation P.57 (1996), *Artificial ears*.

[8] ITU-T Recommendation P.58 (1996), *Head and torso simulator for telephonometry*.

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

[10] ITU-T Recommendation P.810 (1996), *Modulated noise reference unit (MNRU)*.

[11] ITU-T Recommendation P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.

[12] ITU-T Recommendation P.581 (2000), *Use of head and torso simulator (HATS) for hands-free terminal testing*.

[13] ITU *Handbook on Telephonometry*, 2nd edition, Geneva 1992.

1.3 Terms and definitions

This ITU-T Recommendation defines the following terms:

1.3.1 double talk: When near-end and far-end speech occur simultaneously at a given point, typically the terminal under test.

1.3.2 near end: The end of a network connection to which the HFT, whose characteristics are evaluated, is attached.

1.3.3 far end: The end of the network which is opposite to the near end.

1.3.4 syllable clipping or temporal clipping: Loss of speech energy caused by voice/speech activated devices. For echo cancellers, the primary source of temporal clipping is the NLP. In this instance, clipping does not refer to amplitude limiting.

1.3.5 third-party listening test: A listening-only subjective test (see ITU-T Recommendation P.800) in which the listener hears as an "ear witness" the acoustical recordings of the connection under evaluation. In conventional listening-only tests, the listener is positioned at one end of the connection under study.

1.3.6 conversation test: A subjective test in which two participants have a conversation, as described in Annex A/P.800 and in the *Handbook on Telephonometry*.

1.3.7 double talk test: A subjective test in which the participants are forced to talk simultaneously while simultaneously listening for impairments (e.g. echo).

1.3.8 untrained subject: See 3.3.1.

1.3.9 experienced subject: See 3.3.2.

1.3.10 experts: See 3.3.3.

1.3.11 ear signal: Signal recorded in the ear canal of a listener's ear.

1.4 Abbreviations

This ITU-T Recommendation uses the following abbreviations:

ACR Absolute Category Rating

DCR Degradation Category Rating

DMOS Degradation Mean Opinion Score

GAT Group Audio Terminal

HATS Head And Torso Simulator (Recommendation P.58)

HFT Hands-free Terminal

LRGP Loudness Rating Guard Position

MNRU Modulated Noise Reference Unit

MOS Mean Opinion Score

MRP Mouth Reference Point

NLP Non-Linear Processor

2 Overview on test procedures

The test procedures suitable for the assessment of speech quality performance of hands-free terminals can be classified into three categories:

1) Conversational tests (see clause 4).

2) Double talk tests (see clause 5).

3) Third-party listening tests (see clause 6).

NOTE – Although headsets allow a "hands-free" operation and may be tested in general using the same procedures, they are not covered by this Recommendation.

In order to give guidance for the selection of the appropriate test procedure, information about advantages and/or disadvantages of a specific test procedure is listed in Table 1.

Table 1/P.832 – Advantages and disadvantages of different test procedures

	Advantages	Disadvantages
Conversational tests	• Very close to a real conversation • Preparation time is relatively short (compared to third-party listening tests)	• Subjects tend to have different behaviour in a conversation (due to culture, personality, etc.) which creates more response variability in assessing speech quality aspects • Since subjects have to concentrate on both running the conversation and taking care of the quality performance, they may be less sensitive to performance or quality • Devices under test and simulation tools must be available at the testing lab and must run in real time
Double talk tests	• Preparation time is relatively short (compared to third-party listening tests) • Evaluation of double talk capability in more detail than in conversational tests • Due to standardized dialogue structures, individual behaviour depending on culture and/or personality affecting double talk is reduced	• Subjects have to concentrate of both reading their text and taking care of the quality performance • Devices under test and simulation tools must be available at the testing lab and must run in real time

Table 1/P.832 – Advantages and disadvantages of different test procedures *(concluded)*

	Advantages	Disadvantages
Third-party listening tests	• Evaluation of specific speech quality parameters • Processing and assessment of offline simulations • Speech processing is reproducible under the same test conditions • Efficient listening test management in listening labs (e.g. 6 or 8 persons in a listening group) • Application of standardized test and evaluation procedures	• Subjects are not actively involved in the conversation • Speech processing requires measurement and recording equipment • Preparation of a third-party listening test is more time consuming than conversational or double talk test

3 General considerations

Unless otherwise noted, the general considerations described in this clause apply to each of the test methods described in clauses 4-6.

3.1 Hands-free parameters to evaluate

The capability for evaluation of a specific set of speech quality aspects requires different levels of experience of the subjects that conduct a dedicated test procedure. Table 2 provides the parameters to be evaluated by different levels of experience of the subjects (see 3.3.1 and 3.3.2).

Table 2/P.832 – Parameters to be evaluated by different levels of experience of the subjects

	Parameter	Types of subjects
Conversational tests	• Overall quality • Difficulties in talking or hearing • Dialogue capability • Speech sound quality • Transmission of background noise • Variations of loudness during single and/or double talk • Impairments caused by echoes during single and/or double talk	**Untrained subjects:** Evaluation of overall impressions (see clause 4), typically only a few (overall) parameters can be judged at one time. **Experienced subjects:** More detailed evaluation (see clause 4).
Double talk tests	• Overall speech quality • Speech sound quality • Dialogue capability • Transmission of background noise • Completeness of speech transmission • Variations of loudness during single and/or double talk • Impairments caused by echoes during single and/or double talk	**Untrained subjects:** Ratings typical for the average telephone user. **Experienced subjects:** Detailed information about individual degradations.

Table 2/P.832 – Parameters to be evaluated by different levels of experience of the subjects *(concluded)*

	Parameter	Types of subjects
Third-party listening tests	• Dialogue capability • Completeness of speech transmission • Speech sound quality • Variations of loudness during single (and double) talk • Transmission of background noise • Impairments caused by speech gaps • Impairments caused by echoes • Impairments caused by loudness variations	**Untrained and experienced subjects**
	• Impairments caused by level differences between single talk and double talk • Variations of loudness during double talk	**Experienced listeners only**
	• Impairments caused by switching characteristics	

3.2 General considerations about test equipment and calibration

Selection of test equipment, and calibration of the equipment, will depend on the objectives of the test and the application of the hands-free terminal under test. It is, therefore, difficult to provide comprehensive guidance on these issues. However, those conducting subjective evaluations of HFTs should pay particular attention to the following for both near end and far end:

– room conditions (room acoustics);

– delay or other impairments in the "network" (between the devices under test);

– background noise conditions (level, position and kind of sources, e.g. car noise, voice babble, etc.);

– speech characteristics (e.g. level, spectrum);

– speech material for the third-party listening test and double talk test;

– far-end terminal (e.g. handset or hands-free terminal);

– physical arrangement of the set in its intended use condition;

– typical settings of the user.

3.3 Selection of subjects

In general ITU-T Recommendation P.800 should be taken into account for the selection of test subjects.

Some care should be taken when selecting subjects for evaluation of HFTs. As with other speech signal processing equipment, some potential subjects will be more experienced than others. It is recognized that experience with HFTs is a continuum ranging from those who are completely unfamiliar with HFT operation ("non-experts") to those who are thoroughly conversant in the operation and maintenance of HFTs ("experts"), such as HFT designers. However, it is convenient to refer to two parts of this continuum: untrained subjects and experienced subjects.

3.3.1 untrained subjects: Untrained subjects are accustomed to daily use of a telephone. However, they are neither experienced in subjective testing nor are they experts in technical implementations of HFTs. Ideally, they have no specific knowledge about the device that they will be evaluating.

3.3.2 experienced subjects: Experienced subjects (for the purpose of HFT evaluation) are experienced in subjective testing, but do not include individuals who routinely conduct subjective evaluations. Experienced subjects are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are able to describe their subjective impressions in detail. However, experienced subjects neither have a background in technical implementations of HFTs nor do they have detailed knowledge of the influence of particular HFT implementations on subjective quality.

3.3.3 experts: Experts (for the purpose of HFT evaluation) are experienced in subjective testing. Experts are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are able to describe their subjective impressions in detail. They have a background in technical implementations of HFTs and do have detailed knowledge of the influence of particular HFT implementations on subjective quality.

Experts may be used in order to optimize the performance of a hands-free terminal in a very efficient way. Experts may be used for all types of tests. Care should be taken in case only experts are used in a test since they may focus on parameters not of significance for the average user while missing other parameters average users may find significant. Typically the expert's judgement is validated by untrained subjects representing the average user group the set is intended to be used for.

Since typically only a few experts are available, experts tests are conducted mostly for design optimizations during the development process. Due to the low number of experts available the results are normally only of poor statistical significance. Such experts tests are not referenced further in this Recommendation.

4 Conversational test procedure

4.1 Purpose

A conversation test involves two parties conversing over a connection, and depending on the purpose of the test, either experienced or untrained subjects can be used. Such tests can be useful to both manufacturers and customers, and are an important assessment tool because they provide the closest simulation of real telephone interactions between subscribers. The purpose of conversational testing will tend to be different depending on whether experienced or untrained subjects are used.

Untrained subjects are used when it is important to get an indication of how the general telephone-using population would rate the overall quality and difficulty in using the connection with the hands-free telephone. This can be used to give a "global" evaluation of the performance in a range of connections. However, untrained subjects are unable to describe and identify accurately the types of degradation associated with the hands-free telephone.

Experienced subjects are therefore used in the following situations where it is necessary to obtain information about the subjective effects of individual degradations:

1) Diagnosis of hands-free telephone problems.
2) Identification of individual hands-free telephone parameters such as quality of background noise transmission or convergence time (if echo cancellers are included).
3) Establishment of sensible hands-free telephone parameter values.
4) To help choose suitable conditions for inclusion in a test to be performed by untrained subjects.

NOTE – Experienced subjects can judge the overall opinion/quality and difficulties in telephoning as well, but special care should be taken in analysing these results, because experienced subjects typically do not represent the average population.

4.1.1 Benefits

The benefit of conversational testing is that it is the only way of realistically assessing the combined subjective effect of all the parameters affecting conversational quality. In particular, effects such as level variations, echo and double talk can have a marked effect on hands-free telephone performance.

4.1.2 Drawbacks

The drawback of conversational testing is that the conduction of the test is time consuming. In general only a limited set of parameters can be evaluated. Also, the number of conditions that can be realistically tested in one experiment is limited, because of the time required for typical conversations. It can also be quite complex to set up initially.

4.2 Test parameters

The test parameters of conversational tests procedures will tend to be different depending on whether experienced or untrained subjects are used. The parameters which have been successfully evaluated with conversational tests are given below:

The untrained subjects should be asked the following questions:

What is your opinion of the connection you have just been using?

Excellent

Good

Fair

Poor

Bad

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

Yes

No

In case subjects express difficulties they may be asked about kind and nature of the difficulties perceived. Care should be taken in order not to influence the subjects during the questionnaire to avoid statements which are likely to be influenced by the tester.

Further details on these scales are given in ITU-T Recommendation P.800 and the ITU-T *Handbook on Telephonometry*.

Additional questions which possibly might be asked are given below:

How would you assess the dialogue capability?

or alternatively:

How would you assess your ability to converse back and forth during the conversation?

Excellent

Good

Fair

Poor

Bad

How would you assess the sound quality of the other person's voice?

Excellent

Good

Fair

Poor

Bad

Questions possible in addition:

If you heard echo, how annoying was it?

(The test lab should make sure that the test subjects really understand what is meant by "echo".)

Not noticeable

Noticeable, but not annoying

Slightly annoying

Annoying

Very annoying

If there was noise on the connection, how annoying was it?

Not noticeable

Noticeable, but not annoying

Slightly annoying

Annoying

Very annoying

How would you assess the transmission quality of the background noise?

(Open answering form.)

NOTE – If a more specific evaluation is required, it is up to the test lab to define the appropriate questions and the corresponding scales.

The experienced subjects may be asked more detailed questions in addition. Care should be taken when mapping the results achieved with experienced test subjects to the ones achieved with untrained test subjects.

4.3 Test set-up

The general test design, set-up and procedure for full conversation tests are described in ITU-T Recommendation P.800 and the *Handbook on Telephonometry*, which should be consulted for further detail. Some particular considerations for designing full hands-free telephone conversation tests are listed in the clauses that follow.

The test should be designed with a spread of good and bad conditions if possible to ensure that the range of the opinion scale is used as fully as possible.

Compromises have to be made on test duration and the choice of conditions.

Environmental conditions should be chosen to adequately exercise the hands-free telephone and cover the situations where it is likely to be deployed. Test conditions of particular relevance to hands-free telephone testing include:

Environmental conditions:

– room characteristics suitable for the devices under test, e.g.:
  - hands-free telephones for private/business use;
  - conference systems;
  - mobile systems;
  - internet telephony.

Background noise:

– level of noise;

– type of noise (car, babble, etc.);

– dynamic range.

NOTE – Information on room acoustics for the test of hands-free terminals are available in ITU-T Recommendation P.340.

4.4 Description of test procedure

In addition to the descriptions for full conversation tests in ITU-T Recommendation P.800 and ITU-T *Handbook on Telephonometry*, the following considerations should be taken into account.

The conversational task should be designed so that the use of hands-free telephones during the test situation is meaningful. In general, the main considerations for choice of task are to ensure that it leads to a clear conclusion of conversation, that the conversation is not too one-sided, and that a reasonable range of vocabulary is used. In addition, it is important for hands-free telephone testing that the task leads to conversations with a realistic number of double talk situations. Further study would be required to determine a number for this, and it almost certainly varies for different languages and cultures.

Different conversational tasks have been tried by different Administrations, including one where subjects are asked to reach an agreement on an order of preference for a set of picture postcards as described in ITU-T *Handbook on Telephonometry*. Other tasks have also been tried.

In the so-called "Kandinsky test", the subjects are asked to describe to their partner the position of a set of numbers on a picture. Both subjects have similar pictures, but with some of the numbers in different positions. It is recommended that the picture should be designed for the task and that both the picture and the numbers are easy to describe. This can be achieved by using pictures consisting of coloured, geometrical figures (e.g. Kandinsky or others).

In the so-called "short conversational tests", the test subjects are given a task to be conducted at the telephone similar to a daily life situation. The order of a specific pizza at a pizza service or to find a specific railway connection are examples of typical tasks. Since the tests are relatively short, care should be taken in order not to get misled by overestimating the impact of impairments of non-linear and/or time variant systems occurring infrequently during the conversation.

Many conversational tests have been carried out successfully with observers (operators) present in the tests room together with the subjects. It is his task to register and list all comments which subjects mention during or after test. This can be useful for further analysis. Instead of the observers, a video recording might be used.

The ITU-T *Handbook on Telephonometry* also gives some guidance on "simplified conversation tests", where short cuts are suggested to reduce the time taken or to increase the number of treatments in one experiment. Some work has been done with a variation on simplified tests, where subjects are asked to rate a number of individual degradations, after they have given their opinions on quality and difficulty.

4.5 Reference conditions

Reference conditions should be included so that tests performed on different hands-free telephones at different times and by different test labs may be compared. This may be useful, especially for untrained subjects before starting the conversational test, to ensure that all have at least some kind of comparable anchoring.

Such reference conditions may include test set-ups which demonstrate at least the major possible quality degradations with hands-free telephones, such as echo disturbances, level variations and/or switching. As one reference condition a handset-to-handset conversation might be integrated.

5 Double talk test procedure

5.1 Purpose

Comparable to the conversational test, the double talk test involves two parties and, depending on the purpose of the test, either experienced or untrained subjects can be used. These double talk tests are an important evaluation tool because they assess the transmission quality during periods of double talk in detail. Conversational tests clearly pointed out that the double talk performance highly influences the naturalness of a conversation.

The purpose of double talk testing will tend to be different depending on whether experienced or untrained subjects are used.

Untrained subjects are used when it is important to get an indication of how the general telephone-using population would rate the double talk performance with the hands-free telephone. The test procedure is sensitive enough that untrained subjects can assess the relevant parameters even during sophisticated double talk situation.

Experienced subjects are used in the situations where it is necessary to obtain information about the subjective effects of individual degradations:

Diagnosis, parameter identification, parameter value selection, choice of test conditions.

Because of the short duration for single test runs, this specific double talk test is efficient to determine the influence of single subjectively perceived parameters.

5.1.1 Benefits

It is known that the double talk performance highly influences the naturalness of a conversation and consequently the overall quality rating. The benefit of this double talk testing method is that it is designed specially for the quality assessment during the double talk periods and that the test duration is very short. Therefore double talk tests are very efficient to evaluate this very important quality aspect.

5.1.2 Drawbacks

The drawback of double talk tests is that the testing method is more artificial compared to conversational tests because subjects are asked to read a prepared text. Even if the text should be very simple and consists of short, meaningful sentences, the subjects have to concentrate in a different way compared to a free conversation.

5.2 Test parameters

The test parameters of double talk tests will tend to be different depending on whether experienced or untrained subjects are used. In these tests one subject is talking continuously while the other is interrupting (for more detailed information, see 5.4).

This gives the possibility to ask different parameters for both subjects during the test. The following table highlights the parameters which typically determine the double talk performance and were therefore used and successfully evaluated.

Experienced test subjects can be selected to choose test conditions and identify parameters. Experienced as well as untrained subjects can be chosen for parameter value selection and diagnostic purposes.

Continuous talker	Interrupting talker
Double talk capability	Double talk capability
Completeness of speech transmission	Completeness of speech transmission
Loudness during double talk	Loudness during double talk
	Loudness variation single talk/double talk
Echo	Echo
Echo variation single talk/double talk	
Sound quality	Sound quality single talk/double talk
Transmission of background noise	Transmission of background noise

NOTE 1 – Experienced subjects can judge the given parameters for untrained subjects as well, but special care should be taken in analysing these results, because experienced subjects typically do not represent the average population.

The subjects can be asked the following questions:

How would you assess the dialogue capability?

Or alternatively:

How would you assess your ability to converse back and forth during the conversation?

Excellent

Good

Fair

Poor

Bad

How would you assess the completeness of speech transmission during the double talk period?

Excellent – each word is transmitted/intelligible

Good

Fair

Poor

Bad – whole sentences were unintelligible

NOTE 2 – Although the question is multidimensional, test results have shown that test subjects do not have problems with this question. Another possibility could be to ask for speech gaps in combination with a DCR scale. This combination however was not tested yet.

Was your partner's speech ever "cut off" or "clipped"?

Yes

No

How would you assess the loudness during the double talk period?

Much louder than preferred

Louder than preferred

Preferred

Quieter than preferred

Much quieter than preferred

If you heard echo during the double talk period, how annoying was it?

Not noticeable

Noticeable, but not annoying

Slightly annoying

Annoying

Very annoying

How would you assess the sound quality of the other person's voice during the double talk period?

Excellent

Good

Fair

Poor

Bad

In addition the following questions can be asked depending on the task during the tests: (e.g. with an open answering form or any appropriate scale):

Subjects who are "interrupting":

Please compare the (speech) sound quality during single and double talk periods.

How would you assess the transmission quality of background noise during the double talk period?

(Open answering form.)

In addition two different questions can be asked depending on the task during the tests: (e.g. with an open answering form or any appropriate scale):

Subjects who are "talking continuously":

Please compare the echo during single and double talk periods.

Subjects who are "interrupting":

Please compare the loudness during single and double talk periods.

5.3 Test set-up

The general test set-up and environmental conditions for the double talk test are the same as for the conversational tests (see 4.3). Due to the very short duration of a single test run compared to full conversational tests, more conditions can be included than in conversational tests.

5.4 Description of test procedure

During this double talk test, two subjects take part. They are instructed to double talk. The procedure was adapted in the following way:

Both subjects have a written text in front of them. The text itself differs slightly for both. Subject 1 (who is talking continuously) starts reading the text. It consists of simple, short and meaningful sentences. Subject 2 (who has to double talk) can follow the text (without reading), because the beginning of both texts are identical. After some sentences subject 1 skips a passage, which is not given in his text but in the text of subject 2. Subject 1 is reading his complete text which he has in front of him. Thus he cannot compare both texts.

The task of subject 2 is to supplement those sentences which were skipped. He is instructed to start reading at once, if he realizes any missing passage between his own text and the text which subject 1 reads. Correspondingly, subject 1 is instructed not to interrupt reading if subject 2 starts to double talk. The test can be repeated as often as necessary to ensure that both subjects make up their mind about the subjective parameter, which they shall judge during the test.

Both subjects are in a different situation during this specific double talk test. Subject 1, who is continuously reading his text, can only listen to the double talk sequence, if subject 2 interrupts him. It is not possible for him to listen to subject 2 except from this double talk period. Consequently he cannot compare the double talk sequence to a single talk period. On the other hand subject 2 (the one who has to double talk) is in a completely different listening situation. First he listens to subject 1, then he has to double talk. After he finishes, he listens again to the voice of subject 1. Thus he can compare the transmission quality before, during and after double talk. This gives the possibility to ask different parameters for both subjects during the test.

NOTE – Besides the procedure described above, the interruption task also lends itself to a "performance measure" approach. For example, the continuous talker can be given the task of detecting and identifying an interrupting word presented at specific points relative to the text being read by the continuous talker. (Other approaches using recorded conversations and a listening only task can provide even more reliable data.) Performance measures have the advantage of being typically more "objective" than rating scale data. These measurements typically take the form of per cent correct detections and/or per cent correct word identification. One would generally expect a positive correlation between a performance measure and a corresponding rating scale measure. The lack of a positive correlation would suggest further investigation is warranted.

5.5 Reference conditions

Reference conditions should be included so that tests performed on different hands-free telephones at different times and by different test labs may be compared. This may be useful especially for untrained subjects before starting the conversational test, to ensure that all have at least some kind of comparable anchoring.

Such reference conditions may include test set-ups which demonstrate at least the major possible quality degradations with hands-free telephones, such as echo disturbances, level variations and/or switching. As one reference condition, a handset-to-handset condition might be integrated.

6 Third-party listening test procedure

6.1 Purpose

The principle of a third-party listening test procedure is to record specially designed speech material in advance to play it back to the subjects during the test session for evaluation. This test procedure is designed to evaluate and compare the individual performance parameters of different hands-free telephones, different algorithm implementations or different measurement conditions in one test.

Subjects judge the quality of conversational recordings made using a pair of correctly equalized HATS and reproduced by correctly equalized headphones, as third-party listeners. Third-party listener means the subjects are observers of a conversation, standing in the position beside the near-end speaker. The test is applicable to situations where the recording procedure needs to reproduce the listening situation as realistically as possible.

The test may be performed with either untrained or experienced subjects. The purpose of the tests may be diagnostics, parameter identification and parameter value selection.

The test can also be used for the generation of a database of processed speech samples of different hands-free telephones. Such a database may be used to perform comparisons against new implementations.

6.1.1 Benefits

The *measurement conditions* in the set-up can be accurately controlled, and all hands-free telephones may be tested under *identical conditions*. The *numbers of test conditions or hands-free telephones* can easily be increased. If several hands-free telephones, implementations or many environmental conditions are to be included, the procedure is *more economical in terms of time duration* in conducting the tests compared to other tests. Even the number of subjects can easily be increased, and only one set of recordings needs to be made. The simulation of a whole conversation with two artificial head measurement systems allows recordings under *single and double talk conditions*.

The test is well suited to the evaluation of *specific parameters* to give a very detailed and precise description of the achieved transmission quality of the terminals under test because subjects can *concentrate better on these parameters*. The perception of subjectively relevant parameters is in general highly influenced by various parameters like sensitivity, linear and non-linear distortions of terminal equipment, background noise conditions, room characteristics, masking effects and others. The recordings ensure that *a very high degree of realism is reproduced* for third-party listening tests. Subjects judge listening examples, which are recorded at the acoustic interface. Thus all the parameters mentioned above (including masking by the original voice to assess echo disturbances or under double talk condition) are included. Hands-free telephones may be directly judged by *A/B comparisons*. The test is a suitable method for evaluating even small differences between different implementations or different measurement conditions and provides a very efficient procedure to evaluate hands-free telephone differences.

Comparisons between results achieved during conversational tests as compared to third-party listening tests may be found in Annex A/P.340.

6.1.2 Drawbacks

The main disadvantage is – as for listening tests in general – that the results may not be predictive without validation with conversational tests.

The test procedure is *artificial* compared to other tests, where subjects are allowed to talk. Although masking effects and other parameters are considered with this procedure, subjects are asked to listen and judge recordings of unknown speakers. Therefore the masking effects and the naturalness of their own voice speaking is missing. The test procedure cannot therefore be compared to conversational tests or double talk tests. These third-party listening tests are intended to *supplement overall quality evaluations*. They allow detailed parameter investigations, and require *comprehensive preparation.*

Specific conversational related parameters cannot be covered by this test, because this requires complete conversational test with the interaction between subjects.

6.2 Test parameters and scaling

Phone calls using hands-free terminals might be disturbed by many different influences, thus the following different parameters have to be considered:

- overall quality;
- (speech) sound quality;
- impairments caused by speech gaps;
- impairments caused by echoes;
- impairments caused by loudness variations;
- impairments caused by level differences between single talk and double talk;
- switching characteristics during double talk.

In accordance with ITU-T Recommendation P.800, an appropriate five-point category is recommended.

Examples for questions which have been successfully used in third-party listening tests are listed below:

What is your opinion of the connection you have just been using?

Excellent

Good

Fair

Poor

Bad

How would you assess the dialogue capability?

Excellent

Good

Fair

Poor

Bad

How would you assess the sound quality of the other person's voice?

(May be asked for single talk as well as for double talk.)

Excellent

Good

Fair

Poor

Bad

Questions possible in addition:

If you heard echo, how annoying was it?

(May be asked for single talk as well as for double talk.)

(The test lab should make sure that the test subjects really understand what is meant by "echo".)

Not noticeable

Noticeable, but not annoying

Slightly annoying

Annoying

Very annoying

How would you assess the transmission quality of the background noise?

(May be asked for single talk as well as for double talk.)

(Open answering form.)

NOTE 1 – If a more specific evaluation is required, it is up to the test lab to define the appropriate questions and the corresponding scales.

How would you assess the completeness of speech transmission?

(May be asked for single talk as well as for double talk.)

Excellent – each word is transmitted/intelligible

Good

Fair

Poor

Bad – whole sentences were unintelligible

NOTE 2 – Although the question is multidimensional, test results have shown that test subjects do not have problems with this question. Another possibility could be to ask for speech gaps in combination with a DCR scale. This combination however was not tested yet.

How would you assess the loudness during the double talk period?

Much louder as preferred

Louder than preferred

Preferred

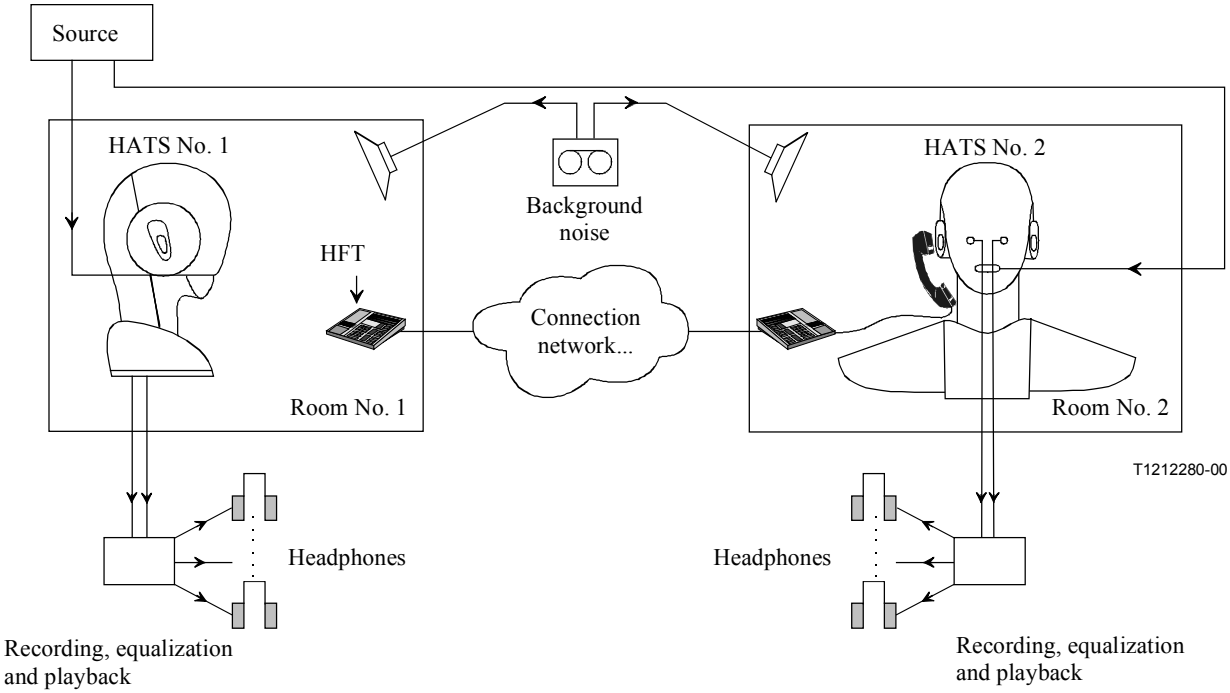
Quieter than preferred (new from P.800)

Much quieter than preferred

6.3 Test set-up and recording parameters

The test material simulates a complete or partial conversation using two HATS according to ITU-T Recommendation P.58, equipped with ITU-T P.58 artificial mouths and ITU-T P.57 artificial ears.

A typical recording set-up is given in Figure 1. It shows a simulated connection with a hands-free telephone on one end and a handset on the other end of the connection. Both subscribers are simulated by HATS according to ITU-T Recommendation P.58. In addition to the P.58 description, the HATS as well as the headphones used for reproduction of the recorded sounds must be properly equalized in order to produce the correct ear signals at the ear of the listener.



NOTE – For reasons of clarity the handset mounting device for HATS No. 2 is not shown.

Figure 1/P.832 – Experimental set-up for recordings of speech material for the listening test using two HATS Exemplary, HATS No. 1 is placed in front of a HFT, while HATS No. 2 is equipped with a handset mounting device for use of handsets

More information about HATS and headphone equalization can be found in ITU-T Recommendation P.581.

All parameters (acoustic environment, speech material and levels, connection network parameters) can be changed for different recording set-ups.

Figures 2 and 3 demonstrate the recordings using HATS as simulated subscribers with a HFT (Figure 2) or handset (Figure 3). All signals and transmission paths which contribute to the ear signals are given.

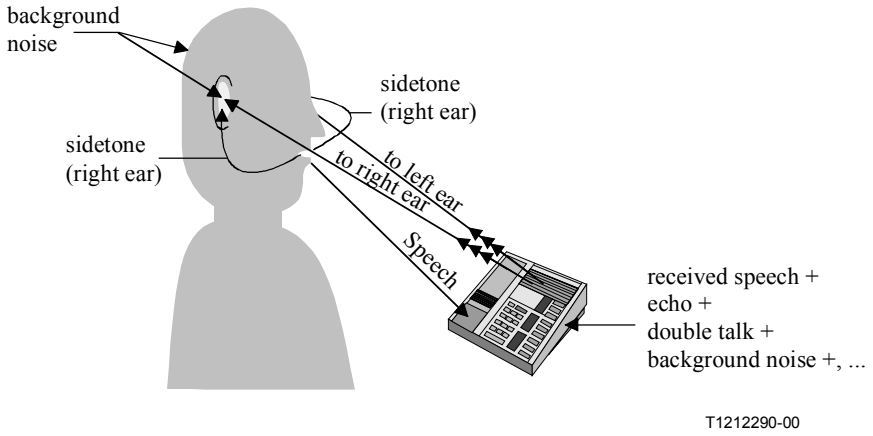


Figure 2/P.832 – Composition of ear signals for a HATS, simulating a hands-free phone user

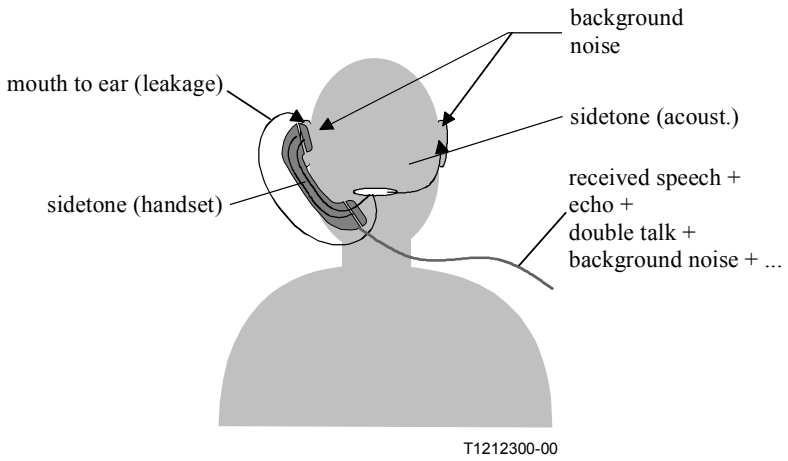


Figure 3/P.832 – Composition of ear signals for a HATS, simulating a handset user

If the HATS simulates the local subscriber with a hands-free phone the system is equipped with ITU-T P.58 artificial mouth and P.57 type 3.3 or 3.4 artificial ears. Figure 2 shows the listening situation in detail. The signals may differ for both ears, depending on room characteristics and the exact location of the hands-free phone relative to the HATS. The speech of the local speaker itself is transmitted to both ears via the acoustical sidetone path. All signals are recorded binaurally, differences between both ear signals depend again on room characteristics and the location of the sources relative to the HATS position in the room.

For handset use the HATS is equipped with ITU-T P.58 artificial mouth and P.57 type 3.3 or 3.4 artificial ears. The handset transmits signals (speech, double talk signals, background noise or echoes) to one ear. The speech of the local speaker itself is transmitted to both ears, but in a different way. The ear which is covered by the handset receives the voice through the leakage between handset and ear and additionally via the sidetone path of the handset. The sidetone path includes electrical sidetone from the handset and acoustical sidetone due to the acoustical coupling between mouth and ear. All these transmission characteristics are pressure force dependent. The other ear, which is not covered by the handset, receives the voice directly from the mouth. The original signal is therefore presented binaurally, but with a significant difference on both ears. The signal, which shall be evaluated, is presented monaurally. The recorded background noise differs again for both ears depending on the acoustical leakage between handset and ear (pressure force dependent), room characteristics and the location of the noise source relative to the HATS.

NOTE 1 – A headset could be used alternatively to the handset.

The artificial mouths are fed from appropriate pre-recorded source material which has typically been recorded and stored on a high-quality digital recording medium, and is then played back. Such a recording allows the preparation and composition of various speech sequences including possible double talk periods if necessary. Equalization before playback ensures that the listening situation is reproduced as closely as possible.

Double talk sequences should be composed in an appropriate way (note that the starting point of double talk can strongly influence the hands-free telephone operation and hence the overall transmission quality, echo disturbances and others). The start of the recordings should be synchronized with the control of other relevant parameters, e.g. control of the hands-free telephones (e.g. if echo cancellers are included) or the start of the background noise (note that in general, the long-term level, level variations vs. time and spectral characteristics of background noise may affect hands-free telephone operation).

For recordings under double talk conditions, care must be taken to distinguish between the near- and far-end speech in the third-party listening test later. From the experience gained with the third-party listening tests, it is recommended that a male and a female voice are used to distinguish between the talkers. One system plays back test sentences of a female voice and the double talk signal is fed from the artificial mouth of the other measurement system using a male voice (or vice versa). If different speakers are used, it is easier for the subjects to concentrate only on the double talk signal during the third-party listening tests.

NOTE 2 – At least one HATS is necessary for the recordings. It should be placed in a suitable location, taking into account room characteristics and background noise conditions. The HATS is always placed at the location where the test material used for the subjective evaluation is recorded. For the simulation of the speaker located at the other end of the connection in principle only an artificial mouth is needed. At this location an artificial mouth according to ITU-T P.51 may be used. If the handset is used at this location it is placed in LRGP. If a hands-free telephone is used at this location it is placed in the appropriate position as defined in the test or in ITU-T Recommendation P.340.

6.4 Description of test procedure

For this third-party listening test procedure prerecorded test material has to be used according to the description in 6.5. One possibility to distinguish between several types of speech sequences is to refer to the included double talk segment:

- Type 1 – short-term double talk.
- Type 2 – long-term double talk.
- Type 3 – without double talk.

The choice of the specific speech sentences is depending on the parameters which have to be evaluated (see Table 3).

It is recommended to use speech material with a male speaker at the far-end side and a female speaker at the near-end side. Under this precondition it is easier to describe the test situation to the subjects. First of all each individual person gets a written instruction for subjects. After that, further instruction has to be given by the conductor of the test session. In addition to that, a short live demonstration in the beginning of each test session gives a precise description of the situation the subjects will be involved in. It could be helpful to have even for this live demonstration a male talker using a real hands-free terminal at the far-end side, and a female talker using the same type of real hands-free terminal at the near-end side, standing just beside the subjects. It is recommended to use for this demonstration hands-free terminals which are not under test.

The test session is held in a studio environment according to [3] (or ITU-T *Handbook on Telephonometry*). For playback of the produced two channel binaural speech material, headphones have to be used in stereo mode. (The test lab should take care that the test subjects are not disturbed by sound emissions of other headphones and environmental noise.) No telephone band filtering is required because of the processing procedure which ensures all signals in the right frequency range. The level relations are identical to that at the artificial head position [about −34 dBPa (A)]; that means the sound pressure level at the artificial head position is exactly reproduced by the test environment. As described in 6.5, the appropriate equalization for the used headphones is required.

The set-up of the experiments is characterized by different blocks, which can be combined to one experiment. In each block of only one parameter (like overall quality for long-term double talk) is evaluated. The different speech samples have to be randomized and the subjects have to assess them absolutely (ACR scale, ITU-T Recommendation P.800). Different types of speech sequences must not be mixed in one block. A short instruction concerning the following parameter and the according practical speech samples have to be given in advance in each block. In total, a duration of 25 minutes for one block should not be exceeded.

One experiment could be composed of several blocks and it must be added that there should be a short break between two blocks even for new instructions for the subjects. In the beginning of each experiment all subjects get a written instruction for careful reading and a verbal explanation as well. Further a live demonstration should be given as described above. The text for this live demonstration should be comparable but not identical to type 2 speech samples. The duration of each experiment is limited to 1.5 hours. The following Table 3 gives an example for the composition of a whole third-party listening test consisting of three experiments and a short overview of the experiment's contents. Each of the experiments consists in that example of three blocks.

Table 3/P.832 – Examples for listening experiments for the evaluation of speech quality aspects of hands-free terminals

Exp. No.	Speech quality aspect parameters	Speech samples	Scale
1	Overall speech quality and sound quality		ACR
	a) Overall speech quality (long time double talk sequences)	Type 1	
	b) Overall speech quality (short time double talk sequences)	Type 2	
	c) (Speech) Sound quality (single talk only)	Type 3	
2	Assessment of different speech impairments		DCR
	a) Impairments caused by speech gaps	Type 1	
	b) Impairments caused by echoes	Type 2	
	c) Impairments caused by loudness variations	Type 1	
3	Impairments with respect to double talk		DCR
	a) Impairments caused by loudness variations	Type 1	
	b) Impairments caused by level differences between single talk and double talk	Type 1	
	c) Impairments caused by switching characteristics	Type 1	

The following roles for the composition of blocks and experiments are recommended:

- Different types of speech samples must not be mixed in one block, therefore, for instance, separate blocks for overall speech quality (long time double talk sequences) and overall speech quality (short time double talk sequences) are necessary.
- If the subjects should be asked for overall speech quality and further parameters in one experiment, it is necessary to ask first of all for the overall speech quality and later on for the other parameter(s); in other words – the overall speech quality have to be the first or the first and the second block of an experiment.
- All other parameters can be asked according to requirements.
- Impairments caused by level differences between single talk and double talk.
- Switching characteristics during double talk.

In accordance with ITU-T Recommendation P.800, a five-point category scale is recommended.

Design of the speech sentences

The artificial mouths of HATS No. 1 and No. 2 are fed with speech sequences for single and double talk. For each type of test different scripts can be created.

For single talk, the far-end or the near-end sentence from the double talk scripts, presented below, may be used.

Scripts for double talk correspond to a conversation between the Local Talker (LCT) and the Far-End Talker (FET). This short conversation consists in a question (one of the talkers) and an answer (the other talker). This can be illustrated by the following example, the Local Talker asking the question, the Far-End Talker answering (in the complete test, the question and the answer are alternatively produced by the Local and the Far-End Talker):

LCT: I suppose I have to take the client file of the new contract of the Company with me

FET: No [*variable silence*] that's not necessary, we have a sufficiently full programme for the first project

The underlined part of the sentences corresponds to a double talk period.

The silence between "no" and the rest of the answer can vary:

– without silence, the sequence corresponds to type 2 test (long double talk);
– with a long silence (more than 5 s), the sequence corresponds to type 3 test (short double talk).

NOTE – In this example (without silence) the duration of the double talk situation is about 2 s.

More information is given in Annex A.

6.5 Reference conditions

Reference conditions can easily be included, because they can be generated offline. These listening examples can be presented with the real recordings during the test. Reference conditions allow results from different labs to be compared, and may include test set-ups under well-defined conditions.

Typical reference conditions should include references for background noise variation, switching loss, switching characteristics and echo loss – depending on the parameter to be evaluated.

NOTE – Reference conditions that include hands-free telephones (comparable to MNRU conditions for codec tests) should be carefully designed to represent typical quality impairments introduced by hands-free telephones. Such conditions should include level variations, echo disturbances or modulated background noise (typically caused by level switching devices or non-linear processes like centre clippers). The same remarks are applicable to reference conditions under double talk conditions. The MNRU, commonly used for codec tests, cannot be recommended, because it does not reproduce the typical quality degradations introduced by hands-free telephones, either under single or double talk conditions.

ANNEX A

Corpus of the source signals

A.1 Size and parameters of the corpus

– Sufficient number of scripts: this point has to be defined according to the corpus size adapted for subjective and objective evaluation.
– In order to not limit the evaluation procedures, the corpus has to be available with several sampling rates (e.g. 48 kHz, 44.1 kHz, 32 kHz, 16 kHz, 8 kHz), with and without filtering (e.g. telephone band filter, etc.), etc.

– For a complete evaluation of hands-free systems, the duration of double talk periods is also an important parameter to consider. The same scripts can be available with several values of double talk duration (e.g. 2 s, 1 s, 500 ms, 250 ms), no double talk with several values of silence duration between the end of the question and the beginning of the answer (e.g. 0 s, 250 ms, 500 ms, 1 s). Note that these different variants can easily be obtained by creating files with appropriate synchronization between the files corresponding to LCT and FET.

A.2 Design of each script

– Each script has to be designed according to the duration of sequences usually used for subjective tests. For example, the total duration has to be around 8 s, with silence periods of 500 ms at the beginning and at the end of each script. For a given script, note that these time values are not constant but depend on the double talk periods (e.g. total duration of 6 s for the case with the maximum double talk duration and 10 s for the one with the maximum silence duration between the end of the question and the beginning of the answer).

– The recording of each sentence has to be done in anechoic conditions with sufficient silence periods at the beginning and at the end of each sentence. This is in order to synchronize in a second step the different versions of each script according to the chosen values double talk or silence periods.

– The coherence of the conversation has to be assured for the maximum double talk duration. In the previous example, the dialogue is coherent because the far-end speaker cut in on the local one just after the words "new contract" are pronounced in the example given in 6.4.

– To optimize the size of the corpus, it is interesting to consider for a given script, two different Local Talkers (e.g. a woman: LCTW, and a man: LCTM) and two different Far End Talkers (e.g. a woman: FETW, and a man: FETM) with the condition that the two different answer sentences (FETW and FETM) are coherent with the two different question sentences (LCTW and LCTM). In this case, four dialogues are available (LCTW with FETW, LCTW with FETM, LCTM with FETM and LCTM with FETW).

SERIES OF ITU-T RECOMMENDATIONS

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

Geneva, 2001