INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

P.85

(06/94)

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

TELEPHONE TRANSMISSION QUALITY

SUBJECTIVE OPINION TESTS

A METHOD FOR SUBJECTIVE PERFORMANCE ASSESSMENT OF THE QUALITY OF SPEECH VOICE OUTPUT DEVICES

ITU-T Recommendation P.85

(Previously “CCITT Recommendation”)

FOREWORD

The ITU-T (Telecommunication Standardization Sector) is a permanent organ of the International Telecommunication Union (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).

ITU-T Recommendation P.85 was prepared by ITU-T Study Group 12 (1993-1996) and was approved under the WTSC Resolution No. 1 procedure on the 21 of June 1994.

NOTE

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

CONTENTS

SUMMARY

Various services providing vocal answers related to telephone directory inquiries, weather forecast, mail order, etc., are now available to PSTN users using vocal servers. As the speech messages are produced by machines, they may suffer from some impairment.

In this Recommendation a method is defined for subjective performance assessment of the quality of speech of voice output devices. This method allows the comparison of several systems between them. It will be useful for system designers and service providers for checking the quality of their products.

This method is of the listening test type. Messages are presented aurally to subjects. The subjects express their opinion on one or more rating scales after having answered specific questions on the information contained in the messages. The results are measures of the perceived quality in several aspects, which makes it possible to compare the effectiveness of different speech synthesis systems.

A METHOD FOR SUBJECTIVE PERFORMANCE ASSESSMENT OF THE QUALITY OF SPEECH VOICE OUTPUT DEVICES

(Geneva, 1994)

1 Scope

Voice servers are now available for Public Switched Telephone Network subscribers. These devices make use either of stored announcements or of synthetic speech. Synthetic speech may be produced from stored segments such as words, syllabes or diphones; it may also be produced by synthesis by rule, e.g. formant synthesis. In all cases of signal processing, such as digital compression of the signal, together with sound processing such as concatenation of segments and variation of pitch, intensity and segment duration, a noticeable impairment of speech quality may occur.

This Recommendation, based on Recommendation P.80 and specific experiments [1], [2], [3], defines a testing method for evaluating the subjective quality of synthetic speech. Some adaptation of the method may be needed, depending on the type of system which is being evaluated.

The method takes into account both the performance and the attitudes of the users. The attitudes are assessed by the use of multiple scales.

The Recommendation covers both overall system performance and the application to specific tasks. Two examples of application are provided in Annex A.

This Recommendation is intended to describe a method for obtaining overall evaluations from users about the acoustic output of speech production devices. Procedures for evaluating specific components of text-to-speech systems (e.g. text transcription into phonetic units, etc.) are currently under study.

2 Assessment method

2.1 General

The recommended methods for assessing telephony speech quality described in Recommendation P.80 and in 2.5 (Opinion tests) of the 2nd edition of *Handbook on Telephonometry* [4] can be applied to the assessment of synthetic speech. The use of multiple opinion scales improves the description of listeners' perception. Since synthetic speech may need some effort to be understood, the test is designed so that the subjects must pay attention to the information contained in messages before expressing their opinions.

2.2 Main features of the recommended method

During a test a number of different voice sources will be presented aurally, so that the subjects' opinions related to a given source may be obtained in relation to other sources. The sources will be synthesis systems as well as reference conditions (this may include natural speech corrupted with some calibrated degradation and known synthesis systems).

Subjects are asked to express their opinion using one or more 5-point opinion scales, as in the Absolute Category Rating (ACR) or Degradation Category Rating (DCR) method of Recommendation P.80. In addition to the overall quality scale, other scales measuring listening effort, voice pleasantness, etc., can be used.

The messages transmitted by the systems should be related to practical applications. In practice different applications will require different test sessions.

Each message is presented twice. During the first presentation subjects answer specific questions on the information contained in a message; then subjects judge the speech quality by expressing their opinion on one or more rating scales during the second presentation.

3 Test preparation

3.1 Speech material

The messages should be long enough so that the subjects have time to reproduce the essential content on the first response sheet and also to give their opinion using the rating scales on the second sheet. A duration of 10 to 30 seconds per message is recommended.

Each message should consist of a fixed part which is specific to the task and a variable part which is different between pairs of presentation. The messages should be designed so that the predictability of the variable part does not differ significantly from one message to another. In Annex A some examples of such messages are given. Other samples with different degrees of difficulty (load of short-time memory) may be used.

3.2 Source conditions

If possible at least five different sources are recommended, depending on the systems to be tested, applications involved and experimental design. Among these sources it is recommended to use at least one natural voice (male or female according to the test systems). The natural voice(s), degraded with a multiplicative noise conforming to Recommendation P.81 (see B.2.3/P.80, "Reference conditions"), should be used as reference. However, research under progress suggests that other degradations may be more suitable to the evaluation of synthetic voices, i.e. T-Reference System [6] or Time and Frequency Warping (TFW) [7].

3.3 Stimulus preparation

This subclause is the same as B.1/P.80 (Source recordings), except that a microphone with a flat frequency response should be used for the recording of the natural voice.

4 Design of experiment

4.1 Subject task

Subjects are given response sheets together with the test instructions. They are requested to use two sheets per message: one sheet is used for reproducing information contained in the message; the other is used for obtaining the subjects' responses on a number of opinion scales.

4.2 Rating scales

The recommended rating scales are:

– overall impression	(type I and type Q questionnaires)
– listening effort – comprehension problems – articulation	(type I questionnaires)
– pronunciation – speaking rate – voice pleasantness	(type Q questionnaires)
– acceptance	(type I and type Q questionnaires)

The wording of the questions and the scaling grades are presented in Annex B.

4.3 Experimental design

4.3.1 Graeco-latin squares (GLs) should be used if the number of source conditions is sufficient, i.e. seven or more. The four factors are: source condition, message, order of presentation, group of subjects.

4.3.2 Within a session, the messages are related to one application only. Similar but different messages should be used for the necessary replications.

4.3.3 When a message has been listened to twice, it shall not be used again.

4.3.4 If all the scales are used, a session will be divided into two blocks, each block corresponding to a type I or type Q questionnaire (see Annex B). If GLs are used, each of the two blocks of a session shall be organized according to two different GLs.

4.3.5 A visit may consist of one or several sessions. Before the main sessions, a training session should be arranged. In the training session, at least six messages should be presented over sources that are sufficiently different to cover the quality range encountered in the test.

4.3.6 If GLs are used, the number of subjects should be at least 4 x GL-dimension (i.e. at least four subjects in each group).

4.3.7 Typical time between two presentations in a pair may be eight seconds, and 20 seconds between pairs, but will depend on the actual message duration.

4.3.8 A visit may last 40 to 60 minutes, including instructions, preliminaries and pauses.

4.3.9 If natural voices are used, one of them should be included into the training session.

4.4 Listening test procedure

4.4.1 *Listening environment* – Same as B.4.1/P.80

4.4.2 *Listening system* – Same as B.4.2/P.80.

All sources should be band-pass filtered in the same way (according to the application, e.g. 300-3400 Hz).

4.4.3 *Listening level* – A target should be that the messages are presented at the preferred level for synthetic speech. If not known, the preferred level for coded speech (79 dB/SPL, –15 dB/Pa, see 2.5.8.1 of the new version of the Handbook on Telephonometry) should be used. If possible one or more test blocks should be presented to the same subjects at two additional levels, one above, one below the preferred level.

4.4.4 *Listeners* – Same as B 4.4/P.80.

4.4.5 *Instructions to subjects* – Annex C gives an example of instructions to subjects; instructions must be given in their written form. They may also be presented verbally, preferably using a tape.

5 Statistical analysis and reporting the results

It is recommended to summarize the opinion scores of the subjects in the form of histograms and/or cumulative distributions for each rating scale.

The comparison of different sources is recommended to be done by plotting the cumulative distributions for each source (one diagram per scale) (see Figure 1).

For the overall quality scale and the listening effort scale it is also possible to calculate the mean opinion scores (MOS) for each source condition and each type of message. An analysis of variance and HSD (Honestly Significant Difference) multiple comparison tests should be made for each rating scale for which MOS values have been calculated.

There is no recommended method for analysing the answers on the information content of the messages. However it may be possible to draw some conclusions if performance (e.g. percentages of correct answers) is noticeably worse for a particular source than for the others.

The results on the acceptance question should be given as percentage values.

The results of the training sessions are not to be used.

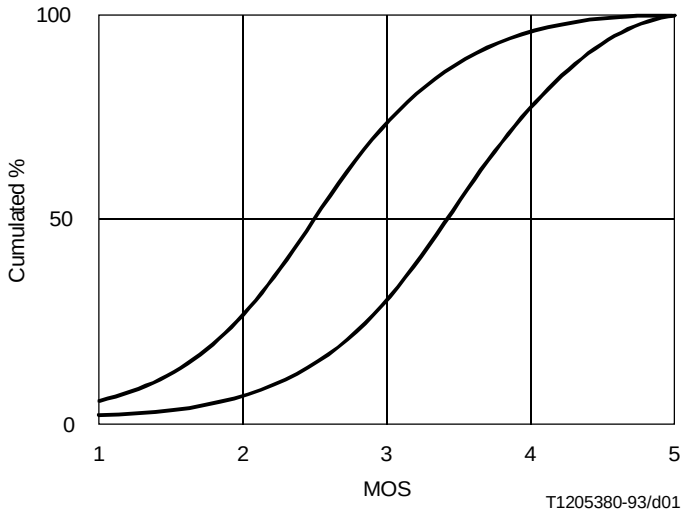

FIGURE 1/P.85

MOS cumulative distributions

6 Other Methods

Sentence-level tests for the assessment of text-to-speech (TTS) systems are especially useful to quantify the overall intelligibility of a synthesiser. Such a test has been designed in the frame of a multi-lingual European project on synthesiser and recogniser assessment (Esprit "SAM" Project No. 2589), the SUS ("Semantically Unpredictable Sentences") test, which has been developed principally for performance evaluation of TTS systems under development [5].

Annex A

Messages

(This annex forms an integral part of this Recommendation)

This annex gives examples of messages. These examples are based on the experiment described in [3].

Two applications were involved in this experiment: mail order shopping (M) and railway traffic information (R). Three messages are given for each application.

M1: Miss Robert, the running shoes colour: white, size: 11, reference: 501-97-52, price: 319 francs, will be delivered to you in 1 week.

M2: Mr. Johnson, the multistandard TV set with remote control, 36 cm screen, reference: 811-61-32, price: 2 492 francs, will be delivered to you in 3 weeks.

M3: Mr. Moore, the electric drill D162, power: 550 watts, 2 speeds, reference: 481-20-30, price: 499 francs, will be delivered to you in 2 weeks.

R1: The train number 9783 from Glasgow will arrive at 9:24, platform number 3, track G.

R2: The train number 7826 to Ipswich will leave at 12:20, platform number 9, track A.

R3: The train number 4320 from Birmingham will arrive at 5:44, platform 2, track C.

Annex B

Response sheets

(This annex forms an integral part of this Recommendation)

The following figures give examples of response sheets. Figures B.1 and B.2 are related to the same applications as in Annex A. See Figures B.3 and B.4.

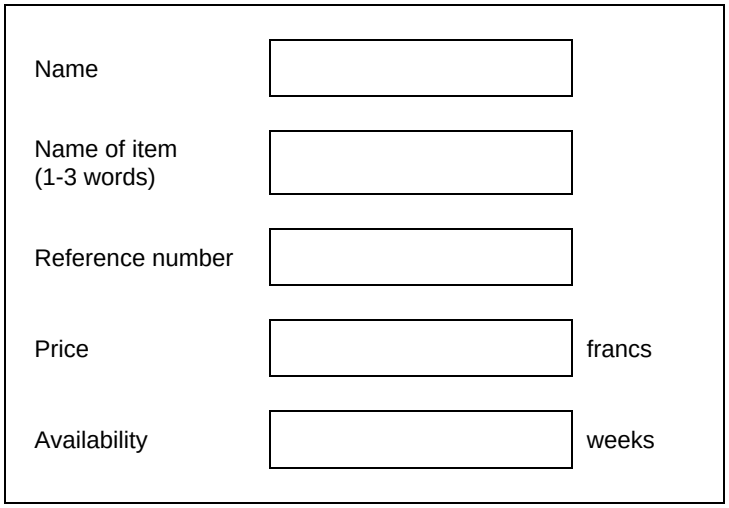

FIGURE B.1/P.85

The five tasks related to a mail order shopping application

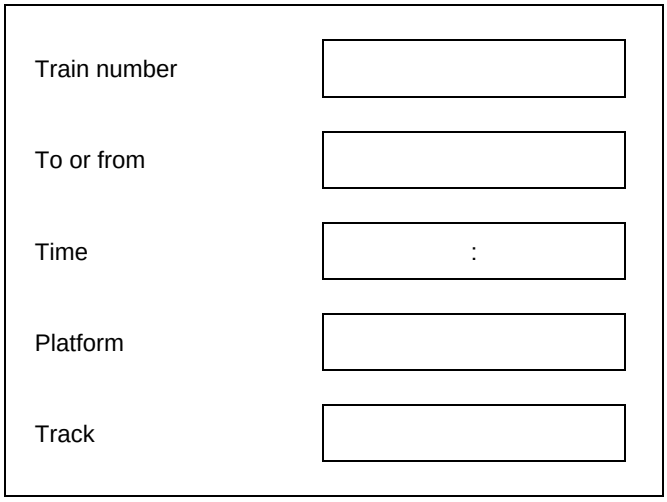

FIGURE B.2/P.85

The five tasks related to a railway traffic information application

Overall impression

How do you rate the quality of the sound of what you have just heard?

- ☐ Excellent
- ☐ Good
- ☐ Fair
- ☐ Poor
- ☐ Bad

Listening effort

How would you describe the effort you were required to make in order to understand the message?

- ☐ Complete relaxation possible; no effort required
- ☐ Attention necessary; no appreciable effort required
- ☐ Moderate effort required
- ☐ Effort required
- ☐ No meaning understood with any feasible effort

Comprehension problems

Did you find certain words hard to understand?

- ☐ Never
- ☐ Rarely
- ☐ Occasionally
- ☐ Often
- ☐ All of the time

Articulation

Were the sounds distinguishable?

- ☐ Yes, very clear
- ☐ Yes, clear enough
- ☐ Fairly clear
- ☐ No, not very clear
- ☐ No, not at all

Acceptance

Do you think that this voice could be used for such an information service by telephone?

- ☐ Yes
- ☐ No

Observations:

FIGURE B.3/P.85

Type I questionnaire in the case when all the scales are used

Overall impression

How do you rate the quality of the sound of what you have just heard?

- ☐ Excellent
- ☐ Good
- ☐ Fair
- ☐ Poor
- ☐ Bad

Pronunciation

Did you notice any anomalies in pronunciation?

- ☐ No
- ☐ Yes, but not annoying
- ☐ Yes, slightly annoying
- ☐ Yes, annoying
- ☐ Yes, very annoying

Speaking rate

The average speed of delivery was:

- ☐ Much faster than preferred
- ☐ Faster than preferred
- ☐ Preferred
- ☐ Slower than preferred
- ☐ Much slower than preferred

Voice pleasantness

How would you describe the voice?

- ☐ Very pleasant
- ☐ Pleasant
- ☐ Fair
- ☐ Unpleasant
- ☐ Very unpleasant

Acceptance

Do you think that this voice could be used for such an information service by telephone?

- ☐ Yes
- ☐ No

Observations:

FIGURE B.4/P.85

Type Q questionnaire in the case when all the scales are used

Annex C

Evaluation of Synthetic Speech: instructions for listeners[1)]

(This annex forms an integral part of this Recommendation)

You are about to participate in an experiment for evaluating various speaking machines.

You will hear two kinds of messages: messages involving mail order shopping and messages involving railway traffic information. The following is an example of each kind of message:

- Mrs Morin, the 50-memory Data-Bank watch with resin bracelet your ordered, reference number 811.19.04, priced at 479 francs, will be delivered in 3 weeks.
- Train number 4119 from New York will arrive at 12:23 at platform 8, track H.

Each message will be repeated twice. During the first presentation, please fill in the boxes, provided for checking your understanding of the information contained in the message. The second time you listen, you will be asked to judge the voice quality of the message by answering a questionnaire consisting of five questions. Some of these questions will change between different parts of the test.

The judgements you are asked to do comprise the following items:

Overall impression – Please try to imagine what your reaction would be if this were an actual telephone message from a mail order house or a request for information from a travel agency.

Listening effort – Your answer should indicate the amount of effort which you were required to make in order to understand the gist of the message and to pick out the information you were asked to reproduce.

Comprehension problems – Please indicate to what extent the content of the message was difficult to understand. This question pertains to all of the words in the message and not only to those you have reproduced.

Articulation – Please evaluate how clear you found the pronunciation (how well you could distinguish the sounds).

Acceptance – Please indicate whether or not you find that the voice you heard would be acceptable for such an automatic answering service by telephone.

Overall impression – Please try to imagine what your reaction would be if this were an actual telephone message from a mail order house or a request for information from a travel agency.

Pronunciation – This question involves possible deviations from natural pronunciation (intonation, rhythm, phrasing).

Speaking rate – Your answer should reflect what your reaction to the speed of delivery would be if this were a real situation.

Voice pleasantness – This question involves your attitude to the voice and whether or not you found it pleasant to listen to.

Acceptance – Please indicate whether or not you find that the voice you heard would be acceptable for such an automatic answering service by telephone.

For each message there are two response sheets. One contains the boxes reserved for answers to specific questions. The other contains a questionnaire with rating scales. (*NB*: you are requested to fill in one sheet at a time and then turn to the next sheet between the presentation of messages; you are not allowed to turn back to any previous sheet.)

The test will start with six practice messages to familiarize you with listening to messages and answering questionnaires. They will provide you with an opportunity to hear examples of systems and voices used in the test and to answer the different types of questions. There will be a break after these six messages to allow you to ask for help if you have any problems.

---

1) In this example, the same two applications as in Annexes A and B are involved, and all the scales are used. The test is divided into 2×2 blocks (2 applications, 2 types of questionnaire)

The test is then divided into two parts separated by a pause. In the first part, you will hear fourteen messages involving mail order shopping. Seven of these are to be evaluated using one type of questions and, following a brief pause, the remaining seven using the other type of questions. In the second part you will hear two blocks of messages concerning railway traffic.

Thank you for participation.

References

– ITU-T Recommendation P.80 *Methods for subjective determination of transmission quality*.

– ITU-T Recommendation P.81 *Modulated noise reference unit (MNRU)*.

Bibliography

[1] CCITT Annex to Report COM XII-R 12 (1986), Subjective assessment of automatic voice answering devices, CSELT (Italy).

[2] CCITT Contribution COM XII-176 (1987), Subjective quality assessment of synthetic speech, Swedish Telecom.

[3] CARTIER (M.), EMERARD (F.), PASCAL (D.), COMBESCURE (P.) and SOUBIGOU (A.): Une méthode d'évaluation multicritère de sorties vocales; application au test de quatre systèmes de synthèse à partir du texte, 19es Journées d'Etude sur la Parole (Société Française d'Acoustique et Association Belge des Acousticiens), Bruxelles, 19-22 mai 1992.

[4] Handbook on Telephonometry, 2nd Edition, ITU (to be published).

[5] BENOÎT (C.), GRICE (M.) and HAZAN (V.): The SUS test: a method for the assessment of text-to-speech synthesis intelligibility (paper submitted for publication in Speech Communication):

– In ESPRIT Project 1541 (SAM), Multilingual Speech Input/Output: Assessment, Methodology and Standardization; Extension Phase Final Report (1 April 1988-28 February 1989); compiled and edited by HARLAND (G.), FOURCIN (A.), BARRY (W.J.) and GRICE (M.): University College London, pp. 344, February 1989.

– A six language test sentence generation software can be provided on request. Contact person: BENOÎT (C.), Institut de la Communication Parlée, Université de Stendhal, B.P. 25X, 38040 Grenoble, France.

[6] Bill Cotton: New Reference Condition For Very Low Bit Rate Coder Evaluation, *Globecom' 92 Conference Record*, Vol 3, pp. 1719-1722, December 6-9, 1992.

[7] ITU-T – Contribution COM 12-18 (1993), An on-going series of subjective experiments to assess speech output from text-to-speech systems, *British Telecom*.