

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.808

(06/2021)

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

Methods for objective and subjective assessment of
speech and video quality

Subjective evaluation of speech quality with a crowdsourcing approach

Recommendation ITU-T P.808

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	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
Methods for objective and subjective assessment of speech and video quality	P.800–P.899
Audiovisual quality in multimedia services	P.900–P.999
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400–P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

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

Recommendation ITU-T P.808

Subjective evaluation of speech quality with a crowdsourcing approach

Summary

Recommendation ITU-T P.808 describes a crowdsourcing approach for conducting subjective evaluations of speech quality. In comparison to laboratory tests, tests using a crowdsourcing approach rely on participants that are connected via an online platform, and whose task is to evaluate speech quality in their own environments, using their own devices. This Recommendation gives guidance on the test material, experimental design, and the procedure for conducting listening tests in the crowd. The methods are to be seen as complementary to laboratory-based evaluations which are described in Recommendations ITU-T P.800 and ITU-T P.835.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.808	2018-06-13	12	11.1002/1000/13625
2.0	ITU-T P.808	2021-06-13	12	11.1002/1000/14665

Keywords

Absolute category rating, comparison category rating, crowdsourcing, degradation category rating, listening test, subjective evaluation, subjective testing.

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

FOREWORD

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

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

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

NOTE

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

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

INTELLECTUAL PROPERTY RIGHTS

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

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2021

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

		Page
1	Scope	1
2	References.....	1
3	Definitions	2
	3.1 Terms defined elsewhere	2
	3.2 Terms defined in this Recommendation.....	2
4	Abbreviations and acronyms	2
5	Conventions	3
6	Crowdsourcing listening-only tests	3
	6.1 Database structure	3
	6.2 Design of experiment	3
	6.3 Listening test procedure	5
	6.4 Data analysis and reporting of results	9
	Annex A – Absolute category rating (ACR) method.....	11
	A.1 Opinion scales	11
	A.2 Stimulus presentation	11
	A.3 Statistical analysis	12
	Annex B – Degradation category rating (DCR) method.....	13
	B.1 Introduction	13
	B.2 Opinion scale	13
	B.3 Stimulus presentation	14
	B.4 Statistical analysis	14
	Annex C – Comparison category rating (CCR) method	15
	C.1 Introduction	15
	C.2 Opinion scales	15
	C.3 Stimulus presentation	16
	C.4 Statistical analysis	16
	Annex D – Evaluating the subjective quality of speech in noise.....	17
	D.1 Introduction	17
	D.2 Reference conditions	17
	D.3 Opinion scales	17
	D.4 Stimulus presentation	19
	D.5 Statistical analysis	19
	Appendix I – Example of job design	20
	I.1 Qualification job.....	20
	I.2 Training job	22
	I.3 Rating job	23
	Bibliography.....	27

Recommendation ITU-T P.808

Subjective evaluation of speech quality with a crowdsourcing approach

1 Scope

This Recommendation contains advice to administrations on conducting subjective tests of speech quality with a crowdsourcing approach. It focuses on listening-only tests including the absolute category rating (ACR), degradation category rating (DCR), comparison category rating (CCR), and methods for evaluating the subjective quality of speech in noise. Other rating tasks as well as conversational tests in the crowd are still under study in ITU-T Study Group 12. The methods described here are to be seen as complementary to the recommended methods in [ITU-T P.800] and [ITU-T P.835]; those methods are carried out in a laboratory environment which is better controlled, whereas the crowdsourcing-based methods described here cover a wider range of realistic listening environments and devices and thus their external validity may be higher.

Crowdsourcing-based methods are not expected to replace laboratory testing, as there are fundamental differences between both methods regarding their conception, the participants and their motivation, as well as technical and environmental factors, as detailed in [b-ITU-T Technical]. As a consequence, the results from crowdsourcing-based methods can be expected to deviate to a certain extent from those of laboratory testing. Depending on the target of the evaluation, the appropriate method has to be selected.

Further guidance on the general approach of crowdsourcing-based testing can be found in [b-ITU-T Technical].

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

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

[ITU-T P.800.2] Recommendation ITU-T P.800.2 (2016), *Mean opinion score interpretation and reporting*.

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

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

[ITU-T P.863] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction*.

[ITU-T P.1401] Recommendation ITU-T P.1401 (2020), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 crowdsourcing [b-ITU-T P.912]: Obtaining the needed service by a large group of people, most probably an on-line community.

3.1.2 question [b-ITU-T P.912]: A single event that requires an answer for a crowdworker. A task contains many questions.

3.1.3 task [b-ITU-T P.912]: Set of actions that a crowdworker needs to perform to complete a subscribed part of the test.

NOTE – 3.1.3 follows terminology presented in [b-Hossfeld].

3.1.4 test [b-ITU-T P.912]: Subjective assessments in a crowdsourcing environment.

NOTE – 3.1.2 follows terminology presented in [b-Hossfeld].

3.1.5 vote [ITU-T P.800.2]: A subject's response to a question in a rating scale for an individual test sample or interaction.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 crowdworker: Person performing a crowdsourcing task.

3.2.2 job: A template for tasks including questions and all the information necessary for a crowdworker to accept and complete that task. A task is an instantiation of a job for a particular crowdworker. An experiment may contain one or more jobs.

3.2.3 job provider: Person or entity who creates a job in a micro-task crowdsourcing platform, also known as requester.

3.2.4 micro-task crowdsourcing: Crowdsourcing simple and small tasks in an open call, to a large and undefined crowd which are usually reimbursed by a monetary reward per each piece of work they perform.

3.2.5 micro-task crowdsourcing platform: A platform which manages the relationship between crowdworkers and job providers including maintaining a dedicated panel of crowdworkers and providing required infrastructure like creating jobs, pool of tasks for crowdworkers, and payment mechanisms.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

CCR Comparison Category Rating

CI Confidence Interval

CMOS Comparison Mean Opinion Score

DCR Degradation Category Rating

DMOS Degradation Mean Opinion Score

GUI Graphical User Interface

MNR Modulated Noise Reference Unit

MOS Mean Opinion Score
N-MO Noise Mean Opinion Score
S-MOS Speech Mean Opinion Score
URL Uniform Resource Locator

5 Conventions

None.

6 Crowdsourcing listening-only tests

Previous experience with crowdsourcing-based evaluation approaches has highlighted several advantages of these new approaches: a very large panel of crowdworkers who can easily and rapidly be requested, potentially from groups which are difficult to recruit for laboratory tests, low costs, and realistic settings of tests. However, although crowdsourcing is promising, it is not intended to replace laboratory tests with standardized methodologies. Speech quality assessment in standardized laboratory environments is well-known, and standardized methods have been established in order to limit bias and give reliable and reproducible results as shown in the ITU-T P.800-series of Recommendations. In fact, crowdsourcing faces several challenges (e.g., conceptual challenges in the test design, reliability of users, incentives and payment schemes to motivate users, hidden influence factors in the uncontrolled environment, and statistical analysis of the results) which are still not completely understood, but will become progressively better controlled due to knowledge gained from previous crowdsourcing tests.

The subsequent clauses address the most relevant of these issues for designing and executing crowdsourcing-based speech quality tests in the listening-only situation. As the complexity of the test situation and the test set-up increases for the conversational case, no recommendations are given for this case at the present state. Regarding listening-tests methods, the current clause describes the database structure, the design of the experiment including the crowdsourcing micro-task platform and test duration, the listening-test procedure, as well as the analysis of the results. Specific considerations regarding different test procedures are given in the annexes.

6.1 Database structure

There is no difference between the preparation of source materials for laboratory tests and that for crowdsourcing tests. Therefore, source recordings and selection of circuit conditions should be prepared as specified in the corresponding clauses in [ITU-T P.800]; i.e., clauses B.1-B.2, D.2.1-D.2.2, and E.2 for ACR, DCR and CCR, respectively. When conducting an ACR subjective test in a super-wideband context, it is recommended to follow the procedure suggested in Appendix II of [ITU-T P.863] when creating the database. Similarly, for evaluation of speech in noise, the dataset should be created by following the corresponding clauses in [ITU-T P.835].

It should be noted that the listening device in a crowdsourcing experiment cannot be assumed as known and identical for each crowdworker. Thus, the preparation of source materials should take the variability in listening devices into account.

6.2 Design of experiment

The same principles as specified in clause A.2 of [ITU-T P.800] should also be followed when applying the selected circuit conditions on the source recordings. Furthermore, due to the conceptual differences between crowdsourcing and laboratory-based experiments [b-ITU-T Technical], the following aspect should also be considered.

6.2.1 Crowdsourcing micro-task platform

One of the following approaches should be adapted when implementing the experiment depending on the purpose of the test:

- using in-built functionalities of the host crowdsourcing platform;
- using the crowdsourcing platform for recruiting crowdworkers, and conducting the study in a separate infrastructure.

Using the in-built functionalities of the host crowdsourcing platform is the recommended method, when the following conditions are fulfilled:

- the crowdsourcing platform provides enough potential participants who meet the conditions specified in clauses 6.3.2, 6.3.3, and 6.3.5;
- the crowdsourcing platform provides audio playback functionalities, or means to implement them;
- the crowdsourcing platform provides means for the job provider to select a group of crowdworkers from previous jobs and give them access to a new job. This can be done by specifying custom qualification requirement(s) for each job, and assigning corresponding qualifications to the selected group of crowdworkers.

If using a separate infrastructure, it is recommended to use a framework to ease moderating the experiment. A detailed comparison between available frameworks can be found in [b-Egger-Lampl].

NOTE 1 – It is assumed that fundamental functionalities of crowdsourcing micro-task platforms like handling payments, jobs with *dynamic contents* (a variable part of a job that changes from task to task e.g., stimuli set to be rated), and statistics like completion times per response are provided.

NOTE 2 – It is suggested to use a platform that provides means to filter crowdworkers based on their long-term performance. Giving access to crowdworkers who showed reliable working habits can decrease uncertainty of the collected data. Statistics like *Task approval rate* (percentage of all tasks performed by the worker that have been approved by the respective job provider), and *number of tasks approved* (number of tasks performed by the worker that have been approved by the respective job provider) or comparable ones are recommended.

6.2.2 Duration of test

The test is limited in size by the maximum length of session possible without fatigue, distraction and possibility of losing the collected ratings. As a typical crowdsourcing micro-task takes a couple of minutes to complete, it is recommended to split an experiment session into a chain of tasks in the rating job. Performing a task from the rating job shall take a couple of minutes (i.e., it should contain 5 to 15 stimuli to be evaluated). However, a crowdworker may perform just one task. As a result, some crowdworkers may not rate the entire set of stimuli available in the database which leads to an error variance caused by individual differences. Therefore, one of the following approaches should be adapted depending to the database structure:

- applying a balanced blocks experimental design as described in [b-ITU-T Handbook]. As a result, crowdworkers should be assigned into groups such that the entire corpus of speech materials is rated by the workers as a whole, but each group rates only a subset of that corpus. Considering the corpus contains t talkers, each spoken s samples, and N degradation conditions then usually s stimulus sets can be created each containing $t \times N$ stimuli (i.e., t stimuli per each condition). As a result, each stimulus set should be evaluated in one task of rating job by a listening panel, i.e., a crowdworker, should be able to take only one task from the rating job;
- motivating crowdworkers to perform multiple tasks from the rating job, and consider individual differences during statistical analyses. Assuming the corpus contains s stimuli, and each task of rating job will include k stimuli, then $\left\lceil \frac{s}{k} \right\rceil$ stimulus sets should be created by randomly selecting stimuli from the dataset. It is recommended to give an extra reward (i.e., bonus) to crowdworkers who perform a sequence of tasks from the rating job ideally

evaluating 50% or more of the entire dataset to reduce the error variance associated with individual differences.

NOTE 1 – A crowdworker can quit at any time, therefore the number of crowdworkers assessing each stimulus should be increased when the number of stimuli presented in one rating job is decreased. Satisfactory correlation between laboratory test and crowdsourcing test were observed in ACR method when 10 stimuli were presented in a rating job with 24 crowdworkers assessing each stimulus (i.e., 96 votes per each condition) [b-Naderi].

NOTE 2 – The number of stimuli presented in one session should be adapted when test methods other than ACR are employed. When evaluating the same number of stimuli in a session, the duration of a DCR/CCR session is about twice the duration of an ACR test session, and about three times when evaluating noise suppressed speech.

6.3 Listening test procedure

6.3.1 Listening session

It is recommended to either create three jobs namely *Qualification Job*, *Training Job* and *Rating Job* (see Figure 1) or one job with three sections from which the *qualification* and *training* are only shown when they are required (hereafter section-based design). A job contains an instruction followed by a list of questions. A question might be static (e.g., 'In what year were you born?') or include dynamic part(s) (e.g., a uniform resource locator (URL) pointing to a stimulus which should be evaluated by the worker for that question). Usually, an identifier is used in the job design to represent the variable part. The job provider shall give a list of values for each identifier when creating that job. As a result, the crowdsourcing platform creates one or more tasks by assigning a value from that list to the identifier.

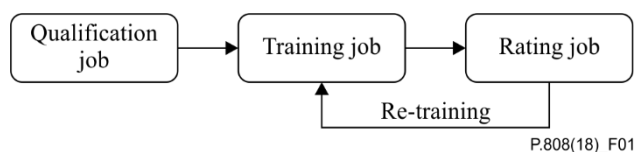


Figure 1 – Workflow of crowdsourcing-based speech quality test

6.3.1.1 Qualification job

Within the qualification job, the purpose of the study should be explained to the crowdworkers, and checked if they are eligible to participate in the study considering the conditions explained in clauses 6.3.3 and 6.3.5. Other evaluations may be considered depending to the aim of the study. It is recommended to use the platform's in-built functionalities to make this job accessible only to crowdworkers who have performed very well in other jobs. Statistics like the *Task approval rate* (e.g., 98% or more) in combination with a sufficiently high *number of tasks approved* are recommended. However, these filters do not guarantee that selected crowdworkers will perform well in the following jobs; therefore, the experimenter must use their own qualification and reliability check mechanisms (see clause 6.3.8) to check the reliability of submitted responses.

Based on the response to this qualification job, a randomly selected group of crowdworkers (who satisfied the prerequisites) should be invited to participate in the experiment i.e., getting access to the training job. The experimenter should consider inviting three to five times of the number of listeners which is expected to evaluate each stimulus.

An exemplary list of items to be included in a qualification job is presented in clause I.1.

NOTE 1 – The experimenter may inform crowdworkers as soon as they get access to the next job.

NOTE 2 – This job should be performed by a large number of crowdworkers to be able to screen for a target group of participants. Therefore, this job should be short and well paid.