

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.808

(06/2018)

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

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

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

Recommendation ITU-T P.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. An Annex describes the details of absolute category rating (ACR) listening quality tests. The method is to be seen as complementary to laboratory-based evaluations which are described in ITU-T P.800.

History

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

Keywords

Absolute category rating, crowdsourcing, 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, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2018

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.....	10
	A.1 Opinion scales	10
	A.2 Stimulus presentation	10
	A.3 Statistical analysis	11
	Appendix I – Example of job design	12
	I.1 Qualification job.....	12
	I.2 Training job	14
	I.3 Rating job	15
	Bibliography.....	19

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 tests and absolute category rating (ACR) tasks. Other rating tasks, such as degradation category rating (DCR) and comparison category rating (CCR), as well as conversational tests in the crowd are still under study in ITU-T Study Group 12. The method described here is to be seen as complementary to the recommended methods in [ITU-T P.800]; the latter methods are carried out in a laboratory environment which is better controlled, whereas the crowdsourcing-based method described here covers 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.78] Recommendation ITU-T P.78 (1996), *Subjective testing method for determination of loudness ratings in accordance with Recommendation P.76.*
- [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.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 (2012), *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, poll 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
DCR	Degradation Category Rating
GUI	Graphical User Interface
MOS	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 the ACR procedure are given in Annex A. Recommendations regarding DCR and CCR are still objects for further study.

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.1-2, D.2.1-2, 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 [ITU-T P.863] Appendix II when creating the database.

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.