INTERNATIONAL TELECOMMUNICATION UNION

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

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/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/.

Recommendation ITU-T P.808 (ex P.CROWD)

Subjective evaluation of speech quality with a crowdsourcing approach

Summary

This Recommendation describes a crowdsourcing approach for conducting subjective evaluations of speech quality. In comparison to laboratory tests, tests using a crowdsourcing approach rely on participants which are connected via an online platform, and whose task is to evaluate speech quality in their own environment, using their own devices. The current 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 ACR listening quality tests. The method is to be seen as complementary to laboratory-based evaluations which are described in [ITU-T P.800].

Keywords

Absolute Category Rating, crowdsourcing, crowdtesting listening test, subjective evaluation, subjective testing

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 SG12. 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, thus their external validity might 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 Technical Report [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 (2014), *Perceptual objective listening quality assessment.*
[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 [ITU-T P.912 Appendix I]: Obtaining the needed service by a large group of people, most probably an on-line community.

3.1.2 test [ITU-T P.912 Appendix I]: Subjective assessments in a crowdsourcing environment.

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

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

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

The 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 [ITU-T P.800-series]. In fact, crowdsourcing faces several challenges (e.g. conceptual challenges in the test design, unreliability of users, but also incentives and payment schemes to motivate users, hidden influence factors in the uncontrolled environment, and statistical analysis of the results to name a few) which are still not completely understood and will be progressively better dominated thanks to the knowledge obtained from previous crowdsourcing tests.

The subsequent sections 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 section 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 an object for further study.

6.1 Database structure

There is no difference between the preparation of source materials for laboratory tests and crowdsourcing test. 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 B1-2, D.2.1-2, and E.2 for ACR, DCR and CCR, respectively. In case of 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 [ITU-T P.800] clause A.2 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 the crowdworkers, and conduct 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 it.
- 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.

In case of 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 an 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 couple of minutes to complete, it is recommended to split an experiment session to a chain of tasks in the rating job. Performing a task from rating job shall take couple of minutes (i.e. it should contain 5 to 15 stimuli to be evaluated). However 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 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 (bonus) to the 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 – A crowdworker can quit at any time, therefore the number of crowdworkers assessing each stimuli 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 when 10 stimuli were presented in a rating job and 24 crowdworkers assessing each stimulus (i.e. 96 votes per each condition) [b-Naderi].

6.3 Listening test procedure

6.3.1 Listening session

It is recommended to create three jobs: *Qualification Job*, *Training Job* and *Rating Job* (cf. Figure 1). Each job contains an instruction followed by a list of questions. A question might be static (e.g. 'In what year were you born?') or includes dynamic part(s) (e.g. a URL pointing to a stimulus which should be evaluated by the worker in 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.
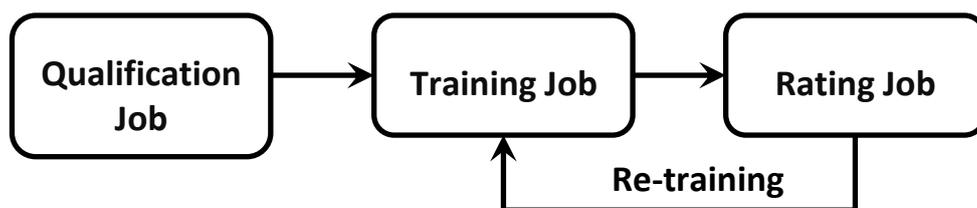
Figure 1- Workflow of crowdsourcing-based speech quality test

6.3.1.1 Qualification Job

Within this 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 just accessible to the 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 approved task* are recommended. It should be noted that those filters do not guarantee that selected crowdworkers will perform well in the following jobs; therefore the experimenter must use their own qualification and gold standard questions (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 Appendix I.1.

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

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

6.3.1.2 Training Job

Within this job, test instructions should be given to the participants as described in clause 6.3.7, followed by a preliminary list of stimuli (i.e. samples). Each crowdworker should listen to the samples and give their opinions on a scale as described in clause 6.3.6. No suggestion should be made to the crowdworkers that the samples include the best or worst condition in the range to be covered. However, in the selection of samples attention should be applied to approximately cover the range from worst to best quality to be expected in the test. The order of presentation of stimuli should be randomized, and each crowdworker should receive the same stimuli for training.

By submitting a response to this job, a temporary access to the rating job should be granted to the crowdworker, i.e. assigning a qualification to that crowdworker. As long as the qualification is valid, the crowdworker can perform tasks from the rating job. Ideally, the access should expire within 60 minutes after granting, which requires the crowdworker to perform the training job again after expiration [b-Polzehl]. In any case, the access should not last for more than 24 hours.

An example of typical training job is given in Appendix I.2.

6.3.1.3   Rating Job

Within this job, first the status of the listening environment, the listening system and level should be evaluated (see clauses 6.3.2, 6.3.3 and 6.3.4). Then, the crowdworker should listen to a set of stimuli and give their opinions as described in clause 6.3.6 and 6.3.7. The number of the stimuli should be decided by the experimenter, following the principles as given in clause 6.2.2. For each stimulus one question should be added to this job. The rating job should contain gold standard questions designed following the requirements of clause 6.3.8. It is recommended to include one gold standard question per each 10 stimuli.

The order of presentation of stimuli in the set should be randomized on runtime for each crowdworker. The experimenter should assign bonuses to crowdworkers who evaluate 50% or more stimuli in the dataset. It is recommended to force the system to download the entire set of stimuli under test in the rating job before the crowdworker can start to rate them, in order to avoid any delay in the rating procedure which might affect the rating.

Appropriate methods may be used within the design of the rating job to make sure that a crowdworker listens to the corresponding stimulus before giving their rating and she/he is providing answers to all questions before being able to submit their response.

In ACR crowdsourcing experiment, each stimulus should be assessed at least by 8 individuals and each circuit condition with 96 votes following recommendations in Appendix II of [ITU-T P.863]. The number of individuals who rate a stimulus should be increased when the number of stimuli presented in one rating job decreases.

An example of typical rating job is given in Appendix I.3.

NOTE 1 – The job provider may consider monitoring the crowdworker's behaviour during the test, including focus time of the browser tab, completion time for each question, and number of time they listen to each stimuli. Those measurements may be used during Data Screening process (see clause 6.4.1).

NOTE 2 – The job provider should warn the crowdworkers in advance that this job needs to download some materials which are free of charge but the downloading file size can lead to some network usage cost.

NOTE 3 – The job provider may consider to merge the training and rating job into one job but with two sections. In that case, the training section should be visible to a worker when it is needed based on the criteria given in the clause 6.3.1.2.

6.3.2   Listening environment

Crowdworkers shall be instructed to perform their task in a quiet and non-distractive environment. They should explicitly be asked at the beginning of each rating job. One of the following approaches can be used to evaluate the listening environment:

- Including a question in which the crowdworkers should record 10 seconds of the environmental noise. Processing the (trimmed) audio files may provide a probability whether an environmental noise level below 50 dB(A) with no dominant peaks in the spectrum is present in the listening environment. Suspected cases of violation will need further evaluation regarding the suitability of the environment.
- Asking the crowdworkers to assess the background noise in their surrounding environment on a five-category intrusiveness scale (Not noticeable, Slightly noticeable, Noticeable but not intrusive, Somewhat intrusive, Very intrusive), see [ITU-T P.835]. Responses given in a surrounding environment which are exceeding a limit (to be determined by the job provider) should be discarded. There is no guarantee, however, that the ratings represent the real loudness of the environment.
- Asking a set of questions in which crowdworkers should compare a pair of speech samples, and give their opinion on which one has better quality. Pairs shall be carefully selected to

differ in quality by a minimum threshold which should also be detectable in the crowdsourcing experiment, and it should be known to the experimenter which stimulus of the pair should have the better quality. When the crowdworker correctly selects the stimuli with better quality in majority of questions, it can be inferred that their surrounding environment and listening system are suitable enough for participating in the study.

The latter is the recommended method.

6.3.3 Listening system

Crowdworkers shall be asked to report required information about their listening system including type of listening device (laptop/desktop loudspeaker, in-ear headphones, over-the-ear headphones) in the qualification job. It is recommended that participants wear a two-eared headphone. However, the experimenter can decide on a different type of listening device, depending on the goal of the test. It should be considered that participants using loudspeakers generally have a smaller discrimination capacity compared to the ones wearing headphones [b- Ribeiro].

The usage of two-eared headphones shall be validated in the beginning of each rating job. A short math exercises with digits panning between left and right in stereo can be used for this purpose (see Appendix I.3).

NOTE – The job provider may ask crowdworkers to take a picture of the headphone using a webcam.

6.3.4 Listening level

Within the instruction of the rating job, the crowdworker shall set the volume of his listening device to a comfortable level when listening to a sample speech file. Afterwards, the crowdworker should not change the listening level when assessing the presented stimuli set in the current rating job. The changes in the listening level may be monitored in case that the crowdsourcing platform provides corresponding means. Responses to a rating task in which the listening level was modified during the test shall be discarded.

6.3.5 Listeners

Crowdworkers taking part in the listening tests are chosen at random from the crowdworkers who responded the qualification job, with the provisos that:

a) they have a normal hearing ability: no crowdworker should exceed a hearing loss of 25 dB at all frequencies up to and including 8 kHz;
b) they are native speakers or presenting a native-level fluency of the language that is used in the spoken material;
c) they have not been directly involved in work connected with assessment of the performance of telephone circuits, or related work such as speech coding;
d) they have not participated in any subjective test whatever for at least for the previous seven days, and not in any listening-opinion test for at least two weeks (not including the current study); and
e) they have never heard the same sentence lists before.

The following demographic distribution of crowdworkers (as proposed in Appendix II of [ITU-T P.863]) may be considered when randomly sampling participants:

f) At least 20% of participants should belong to each of following age groups [15,30), [30,50), 50+;
g) Within each age group, at least 40% of participants should be male and at least 40% should be female.

The experimenter shall ask corresponding questions in the qualification job to be able to evaluate abovementioned conditions. If the available population is unduly restricted, then allowance must be made (except for conditions *a* and *b*) for this fact in drawing conclusions from the results.

Methods for verification of any of the above-mentioned provisos are for further study.

6.3.6   Opinion scales

See clause A.1 for ACR procedure.

6.3.7   Instructions to subjects

The instruction for each job should be given at the beginning of that job. However, a short overall instruction shall be given within the qualification job prior to commencement of the experiment. Besides typical instruction given in the laboratory experiment, following information should be given in the instruction of a crowdsourcing experiment:

* Eligibility requirements and exclusion criteria;
* Estimated time it takes to complete a task (estimated from pilot studies);
* Expectation from participants such as the type of responses that may result in rejected work, listening device and environmental conditions;
* (optional) The identity of the research group. Stating affiliations helps to build trust with the crowdsourcing community.

The experimenter may provide online communication means to answer questions of participants regarding the instruction (e.g. live-chat, or email address). Questions about the procedure or about the meaning of the instructions should be answered, but any technical questions must be met with the response, "We cannot tell you anything about that until the experiment is finished". Short version of instruction can be repeated in the training and the rating job. Specific instructions for the ACR procedure are given in clause A.2.

6.3.8   Gold standard question

The gold standard question (also known as trapping question) is a question that its answer is known to the experimenter. Crowdworkers shall be able to give a correct answer easily when they completely and consciously follow the test instruction. It is recommended that the gold standard questions fulfil the following conditions:

* It should not be easily recognizable unless the crowdworker follow the procedure of the test (no visual and contextual differences with other questions in the rating job)
* The effort of concealing the cheating would be as high as the effort of providing reliable answers
* It makes crowdworkers aware of the importance of their work, in order to motivate

It is recommended to add one or more gold standard questions to the rating job. They should be visually identical to the other questions, but containing a trapping stimulus rather than a normal stimulus from the dataset. A set of trapping stimuli should be created as follows and randomly used:

1) Five stimuli per each speaker from the dataset should be randomly selected, reflecting different degradation conditions.
2) A message should be recorded with a speaker not being part of the speech material, in a same language as the spoken material.
3) A variation of the recorded message (#2) should be appended to the first seconds of each selected stimuli (#1) to create the trapping stimuli set.

The following message should be used as proposed by [b-Naderi] as the message #2: "This is an interruption. We would like to ensure that participants work conscientiously and attentively on our

tasks. Please select the answer *X* to confirm your attention now." where *X* can be any item from the opinion scale (e.g. X = Poor, or Fair in the ACR test). Five variations of this message (one per each opinion scale item) should be created.

The gold standard question(s) should be randomly positioned between the quality assessment questions in the rating job.

More details on designing the audio trapping stimuli can be found in [b-Naderi].

NOTE – In the case of using DCR or CCR procedures, employing at-least one "null pairs" (same stimuli A-A in the pair) in each rating job is recommended.

6.4 Data analysis and reporting of results

6.4.1 Data Screening

Before performing statistical analysis, the submitted responses from the crowdworkers should be a subject of a data screening process. The submitted response to a rating job should be discarded when:

- One or more gold standard questions in the job are answered wrongly.
- The listening system is not used as specified (e.g. changing listening level during session, or using one-eared headphone when two-eared headphone is required)
- The listening environment was not suitable

The experimenter should evaluate the submitted responses against unexpected patterns in ratings (such as no variance, or potential outliers) and unexpected user behaviour in a session (such as listening to a stimulus several times). Univariate outliers can be identified by calculating the standardized scores of entire votes for each stimulus (or condition). Votes with absolute z-score larger than 3.29 should be considered to be potential outliers [b- Tabachnick]. Other outlier detection methods including boxplot (extreme outlier when a rating is beyond an outer fence) might be employed as well. The experimenter may discard a response given in a session when unexpected user behaviour were observed.

All responses submitted by a participant should be removed when his/her responses did not fulfil the abovementioned conditions more than twice.

NOTE – For further discussions on screening mechanisms based on user ratings see [b-Hossfeld].

6.4.2 Statistical analysis
See the Statistical analysis clause in the corresponding Annex depending to employed procedure.

6.4.3 Reporting subjective MOS values

The following information shall be provided when reporting the subjective MOS values obtain through crowdsourcing approach in addition to the information specified in clause 12 of [ITU-T P.800.2].

1. Study: crowdsourcing platform, frameworks (if applicable), payments, requested qualifications (if any), duration of test, number of stimuli in rating job.
2. Subject profiles: number of crowdworkers for each job, for worker who took the rating job: age and gender distribution, used equipment.
3. Data screening process: number of discarded responses and employed criteria.

Annex A

Absolute Category Rating (ACR)

(This annex forms an integral part of this Recommendation.)

A.1 Opinion Scales

In an ACR test, various five-point category-judgement scales may be used depending to the purpose of experiment. The layout and wording of opinion scales, as seen by subjects in experiments, is very important, and should follow the standard arrived at through years of experience. The opinion scales as specified in Annex B.4.5 of [ITU-T P.800] should be adapted to be used in a computer aided system. The scale should be presented in a way that

- Both "term" and "score" are visible for the subject,
- The distance between the points should be equal.

In Figure A.1 an adapted listening-quality scale is presented.

Quality of the speech	Score
◎ Excellent	5
◎ Good	4
◎ Fair	3
◎ Poor	2
◎ Bad	1

Figure A.1 – Adapted opinion scale for assessing the listening quality of speech

The quantity evaluated from the scores (mean listening-quality opinion score, or simply mean opinion score) is represented by the symbol MOS.

NOTE – Other opinion scales presented in Annex B.4.5 of [ITU-T P.800] should be adapted in a same way.

A.2 Stimulus presentation

In each question, one stimulus is presented to the crowdworker and they are asked to indicate their opinion on the given scale. Questions may be presented in a list or different pages. Either standard HTML5 <audio> tag or customized HTML tags can be used for audio playback. It is recommended to avoid providing a volume and seeking controls within audio playback GUI. An example of typical rating question is given in Figure A.2.
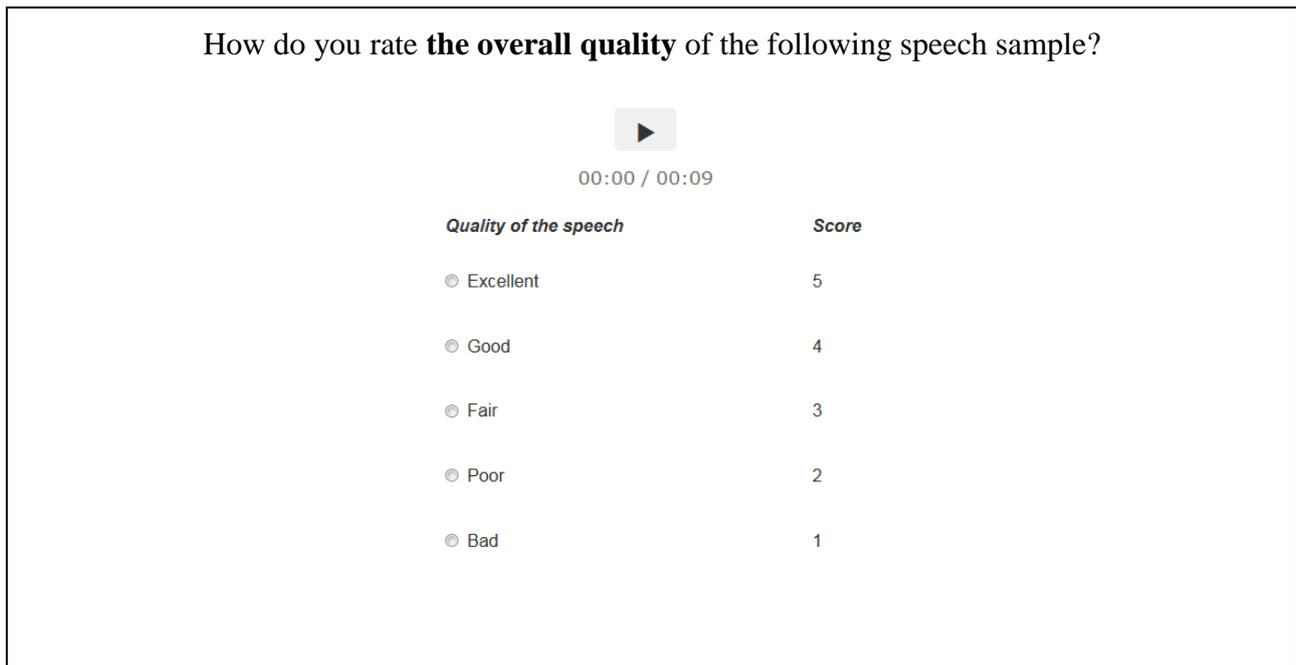
How do you rate **the overall quality** of the following speech sample?

▶

00:00 / 00:09

Quality of the speech	Score
◎ Excellent	5
◎ Good	4
◎ Fair	3
◎ Poor	2
◎ Bad	1

Figure A.2 – Example of presentation of stimulus using a customize playback component.

A.3 Statistical analysis

The numerical mean (over subjects) should be calculated for each stimuli for initial inspection (so that effects such as those due to male and female talker can be seen) and then for each condition.

For each stimuli and condition, MOS values should be accompanied by sufficient information to allow a basic statistical analysis to be performed, for example, the calculation of a confidence interval (see ITU-T P.1401 Appendix III). For any given stimulus and condition, this information comprises the number of votes, the mean of the votes and the standard deviation of the votes. It is recommended to evaluate the confidence intervals of subjective scores when comparing conditions.

Depending to the experiment design, further analysis can be performed using mixed models (individual differences as random effect) or significance tests performed by conventional analysis-of-variance techniques (see [b-ITU-T Handbook]).

Appendix I

Example of Job Design

(This appendix does not form an integral part of this Recommendation.)

In the following sample jobs designed for ACR method based on this Recommendation are given.

NOTE – The terminologies and numbers used should be adapted to the platform and experiment. Further notes are given in brackets within each samples.

I.1 Qualification Job

Following is an example of a qualification job.

NOTE – Crowdworkers who are native speakers or presenting a native-level fluency of the language that is used in the spoken material should get access to this job. In case the crowdsourcing platform do not provide means for giving access to crowdworkers based on their language knowledge, experimenter should design an appropriate language test.

Instruction for speech quality assessment (Part 1 - Qualification)

Introduction

We are looking for crowdworkers who are willing to participate in a **speech quality assessment** experiment. During that test you will listen to **… audio files**, each 6-8 seconds long (two sentences), via your listening device and you will be asked to indicate your opinion quality of each on the following scale:

Quality of the speech	*Score*
○ Excellent	5
○ Good	4
○ Fair	3
○ Poor	2
○ Bad	1

Each of those tasks can be completed in about **… minutes**. There will be a total of **… Tasks** available for each crowdworker. It results to **$ …** compensation including bonuses. Bonuses will be granted based on 1) number of tasks you perform, 2) quality of your work.

Procedure:

1. To get access to the above-mentioned **rating job**, you should first complete this **qualification job**.

---

2. Selected group of crowdworkers will be invited to perform the **training job (… minutes)** in which you will listen to … sample audio files.
3. Then, they get access to the **rating job** and can perform up to … tasks.

Conditions:

-   You must perform the task in a **quiet environment** like at home.
-   You must use **headphones**. Note that, <u>loudspeakers</u> are **not acceptable**.

Thank you for your help in this experiment.

Questions

Please answer the following questions carefully.

1. What is your gender? Male / Female / Other
2. In what year were you born? [TEXT Field/drop-down list] / Or range of age
3. What type of listening devices do you have and able to use now (select all)? [Image & checkbox]
4. When was the last time you participated in a subjective test? [desire: 1 week or later]
5. When was the last time you participated in an audio listening test? [desire: 1 week or later]
6. Have you ever been directly involved in work connected with assessment of the performance of telephone circuits, or related work such as speech coding? [yes/no]
7. I believe, … [radio button]
    -   I have a normal hearing ability.
    -   I have difficulties keeping up with conversations, especially in noisy surroundings (mid hearing loss).
    -   I have difficulty keeping up with conversations when I am not using a hearing aid (moderate hearing loss).
    -   I rely on lip-reading even when I am using hearing aids (severe or profound hearing loss).

Please wear your headphones now.

8.   Please adjust the level of your computer to a comfortable level so that you hear the following audio sample very well.

▶

00:00 / 00:09

[Audio file::calibrate_listening_level]

NOTE: After that, you are not allowed to change the volume anymore. Otherwise your response will be rejected.

[*Next, add your selected hearing test here*]

9. *[self-screening hearing test: digit triplet test]* In each of the following tests, you hear a combination of three digits (for example 1 5 3), spoken with noise in the background. Simply enter the three numbers that you have understood on the answer box below each audio file.

Test	Test	Test	Test	Test
▶	▶	▶	▶	▶
00:00 / 00:04	00:00 / 00:04	00:00 / 00:04	00:00 / 00:04	00:00 / 00:04
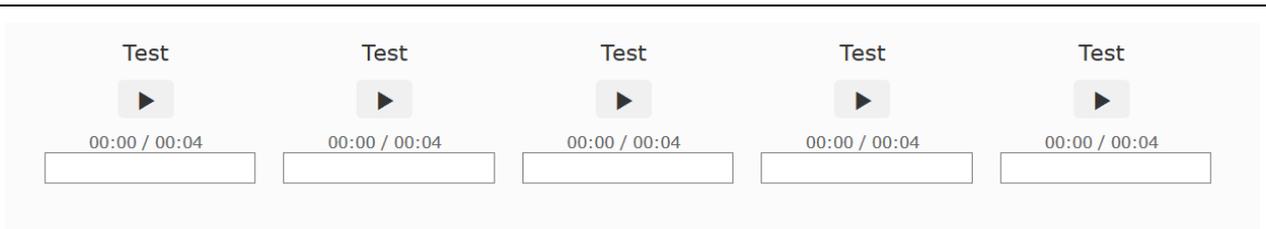
Thank you for your participation. The qualifications will be assigned to selected group of participants in up to next … days.

I.2   Training job

Following is an example of a training job.

Instruction for speech quality assessment - (Part 2 - Training)

Welcome and congratulation!

You have been selected to participate in our speech quality assessment experiment.

This is a training job. By completing this job, you will get a certificate/qualification [platform specific term] which will expire in … hours. Within that time, you can perform the assessment job as much as it is available for you.

Please use … for listening to the audio files.

Setup

1.   Please adjust the level of your computer to a comfortable level so that you hear the following audio sample very well.

▶

00:00 / 00:09

[Audio file::calibrate_listening_level]

NOTE: After adjusting the level, you are not allowed to change the volume anymore. Otherwise, your response will be rejected.

2.   Please listen to the following audio file and type in the answer:

▶

00:00 / 00:09

[Audio file::Math question Y to check two-ear plug]

Training samples:

Please answer the following … questions. For each question, please listen to the audio sample and give your opinion about the quality of the speech you hear on the following scale. Note that the scale will be activated when the speech sample played until the end. In case you hear an interruption message, please follow the instruction given in the message.

There is no right or wrong answer as long as you listen to the audio files and give your opinion.

1. How do you rate **the overall quality** of the following speech sample?

▶

00:00 / 00:09

Quality of the speech	Score
○ Excellent	5
○ Good	4
○ Fair	3
○ Poor	2
○ Bad	1

[...]

5. How would you rate **the overall quality** of the following speech sample?

▶

00:00 / 00:09

Quality of the speech	Score
○ Excellent	5
○ Good	4
○ Fair	3
○ Poor	2
○ Bad	1

Thanks for your participation.

Your qualification will be assigned in … minutes/days. Then you are allowed to perform … jobs.

I.3   Rating job

Instruction for speech quality assessment - (Part 3 - Rating)

Welcome!

This task has two sections:

- **Setup:** Configure your system and validate it by answering to 6 questions
- **Rating:** Listen to … audio files and give your opinion about the quality of the speech you hear.

You can perform as many tasks available to you from this job until your qualification expires (… hours after assigning). When your qualification is expired, you can obtain it again by repeating the training job.

You should follow the below mentioned rules, otherwise your answers will be invalid.

Rules:

- Use a headset, not the loudspeaker: otherwise your response will be rejected
- Perform the task in a quite environment
- Do not change the volume after modifying it in the Setup section.

Payment

The result of this experiment is very important for us and other scientist working in this area. We have methods that analyse the consistency of your answers. We will use these methods to rank the submitted assignments according to quality.

For this experiment we will pay a base reward of $.../HIT for every accepted HIT. We have made available a set of … different HITs. You will receive a bonus of:

- $0.10/HIT (for a total of $0.20/HIT) if you submit all … HITs or
- $0.20/HIT (for a total of $0.30/HIT) if you submit all … HITs and be in the top 20% quality group.

Setup

Please wear your headphones now

1. Please adjust the level of your computer to a comfortable level so that you hear the following audio sample very well.

▶

00:00 / 00:09

[Audio file::calibrate_listening_level]

   NOTE: After adjusting the level, you are not allowed to change the volume anymore. Otherwise, your response will be rejected.

2. Please listen to the following audio file and type in the answer:

▶

00:00 / 00:09

[Audio file::Math question Y to check two-ear plug]

For the following **four questions**, please specify which sample has a better quality compared to the other one from your perspective.

3.1. Which sample has a better quality compared to the other one?

Sample A   Sample B

▶   ▶

00:00 / 00:09   00:00 / 00:09

◯ Quality of **Sample A** is better.
◯ Difference is **not detectable**.
◯ Quality of **Sample B** is better.

[…]

3.4. Which sample has a better quality compared to the other one?

Sample A   Sample B

▶   ▶

00:00 / 00:09   00:00 / 00:09

◯ Quality of **Sample A** is better.
◯ Difference is **not detectable**.
◯ Quality of **Sample B** is better.

Ratings

Please answer the following … questions. For each question, please listen to the audio sample and give your opinion about the quality of the speech you hear on the following scale. Note that the scale will be activated when the speech sample played until the end. In case you hear an interruption message, please follow the instruction given in the message.

There is no right or wrong answer as long as you listen to the audio files and give your opinion.

1. How do you rate **the overall quality** of the following speech sample?

00:00 / 00:09

Quality of the speech	Score
○ Excellent	5
○ Good	4
○ Fair	3
○ Poor	2
○ Bad	1

[…]

X. How do you rate **the overall quality** of the following speech sample?

00:00 / 00:09

Quality of the speech	Score
○ Excellent	5
○ Good	4
○ Fair	3
○ Poor	2
○ Bad	1

Thanks for your participation. Feel free to take more tasks from this job when they are available for you.

Bibliography

[b-Egger-Lampl]   Egger-Lampl, S., Redi, J., Hoßfeld, T., Hirth, M., Möller, S., Naderi, B., Keimel, C. and Saupe, D. (2017), *Crowdsourcing Quality of Experience Experiments*. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, Cham.

[b-Hossfeld]   Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S., Keimel, C. (2014), *Best practices and recommendations for crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing*, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003 Qualinet).

[b-ITU-T Handbook]   ITU-T Handbook (2011), Practical Procedures for Subjective Testing.

[b-ITU-T Technical]   ITU-T Technical Report PSTR-CROW (2018, in press), *Subjective evaluation of quality of media with a crowdsourcing approach.*

[b-Naderi]   Naderi, B. Polzehl, T., Wechsung, I., Köster, F., Möller, S. (2015), *Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm*. 16th Ann. Conf. of the Int. Speech Comm. Assoc.. ISCA, pp. 2799–2803.

[b-Polzehl]   Polzehl, T., Naderi, B., Köster, F., Möller, S. (2015), *Robustness in Speech Quality Assessment and Temporal Training Expiry in Mobile Crowdsourcing Environments*. 16th Ann. Conf. of the Int. Speech Comm. Assoc.. ISCA, pp. 2794–2798.

[b- Ribeiro]   Ribeiro, F., Florêncio, D., Zhang, C. and Seltzer, M. (2011), *Crowdmos: An approach for crowdsourcing mean opinion score studies*. In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, pp. 2416-2419.

[b-Tabachnick]   Tabachnick, B.G. and Fidell, L.S. 2012. *Using Multivariate Statistics*. Pearson Education.

_____