INTERNATIONAL TELECOMMUNICATION UNION

TELECOMMUNICATION STANDARDIZATION SECTOR

STUDY PERIOD 2017-2020

SG12-C.204

STUDY GROUP 12

Original: English

Question(s): 7/12 Geneva, 1-10 May 2018

CONTRIBUTION

Source: Deutsche Telekom AG

Title: Evaluation of the Draft of P.CROWD Recommendation

Purpose: Information

Contact:
Babak Naderi
Telekom Innovation Labs, TU Berlin
Germany
Tel: +49 30 8353 54221
Fax: +49 30 8353 58409
E-mail: babak.naderi@tu-berlin.de

Contact:
Sebastian Möller
Telekom Innovation Labs, TU Berlin
Germany
Tel: +49 30 8353 58465
Fax: +49 30 8353 58409
E-mail: sebastian.moeller@tu-berlin.de

Contact:
Rafael Zequeira Jiménez
Telekom Innovation Labs, TU Berlin
Germany
Tel: +49 30 8353 58336
Fax: +49 30 8353 58409
E-mail: rafael.zequeira@tu-berlin.de

Keywords: ITU-T SG12; speech; crowdsourcing; ACR; Laboratory, P.CROWD

Abstract: This contribution reports result of two crowdsourcing studies conducted based on the drafted document of P.CROWD to compare subjective ratings collected using the proposed crowdsourcing approach with the standard laboratory based tests. Datasets from the ITU-T Rec. P.863 competition were kindly provided including source materials and laboratory-based ratings. From them, two datasets, namely 401 and 501, were chosen and two groups conducted the crowdsourcing studies. Results show strong correlations between laboratory-based ratings and ratings collected through the crowdsourcing approaches following the P.CROWD draft (Spearman’s rank-order correlation of .971 for dataset 401, and .891 for dataset 501).

Introduction

In the following, two studies are described which were conducted to validate the proposed draft of the P.CROWD Recommendation [2]. The recommendation draft explains how to conduct speech quality assessment using crowdsourcing approach first focusing on the ACR.

This contribution organized as following. First, the datasets used for each study are explained, and each crowdsourcing test is described in detail. Later, results of comparisons between crowdsourcing approach and the laboratory experiments are presented.

We invite experts of Q7/12 to consider the results presented in this paper when discussing the draft Recommendation of P.CROWD.

Databases

Access to the pool of the ITU-T Rec. P.863 competition datasets were kindly provided for the aim of this study. From the datasets available in the pool, two were select namely 401, and 501 each providing different language, study design and includes variable types of degradations and degradation combinations and prepared base of the ITU-T Rec. P.800 [4] specification. Table 1 summarize the source materials and laboratory based ratings provided by the corresponding contributors.

Table 1: Selected datasets from the ITU-T Rec. P.863 competition pool, used for evaluation.

	Dataset 401	**Dataset 501**
Title	Psytechnics P.OLQA Proponent test 1 - P.OLQA Full-scale SWB test	SwissQual P.OLQA SWB TrainingDatabase1
Owner	Psytechnics Ltd	SwissQual® License AG, Switzerland.
Date	July 2008	September 2008
Test method	P.800	P.800
Quality Scale	ACR	ACR
Number of Conditions	48	50
Files per condition	24	4
Votes per File	8	24
Votes per Condition	192	96
Listeners	32	24
Design	6 talkers (3m, 3f)	4 talkers (2m, 2f)
Language	British English	German, Swiss pronunciation
# of samples	1152	200
Listened through	Sennheiser HD25-1	Grado SR60

Crowdsourcing test

Two experimenters conducted the crowdsourcing tests each using one of the abovementioned datasets and following the proposed draft of P.CROWD. In the following, each test is described in details.

CS 401

This experiment was conducted using the Amazon Mechanical Turk[1] (MTurk) platform. MTurk is a well-known crowdsourcing platform and provides globally distributed crowdsourcing workforce[2], however their largest population belongs to US and Indian based workers as reported in 2016 [3]. In this study the platform internal infrastructure are used. Jobs were designed using HTML/CSS/JS code. As proposed by the P.CROWD draft multi-step job design were followed.

[1] https://mturk.com/

[2] Our attempt on reaching German speaking crowdworkers in Feb. 2018 was unsuccessful.

Qualification Job: As suggested in P.CROWD platform provided conditions were used to make sure that workers with a history of good performance could only participate in this job (i.e. overall approval rate > 98%, and accepted jobs >500). In addition, workers from US were only able to perform this job and no language screening test were employed. For listening impairment test, the adapted version of digit-triplet test were used. There, five stimuli with signal-to-noise ratio (SNR) of -11.2 dB were used. Workers should listen to each stimulus and type in the three numbers those they hear. They could listen to each stimulus as many times as they want (the number was captured for further analyses). The SNR of -11.2 dB was chosen to reach high true positive rate while previous study suggested to use threshold of -9.3 dB SNR for German, -11.2 dB SNR for Dutch, and -10.5 dB SNR for French digit-triplet test to find normal hearing participants [1].
From 227 participants, 133 workers were not eligible to continue because of the result of hearing test, and 7 other workers because of other reasons like self-reported hearing impairment and inadequate listening devices. From 187 remaining ones, 100 were randomly select (following the proposed distribution of gender and age in the draft text) and given access to the next job.

Training job: As the temporal qualification condition is not supported by the crowdsourcing platform, it was decided to merge the training job and rating job into one single job and implement a temporal qualification using Cookies. As a result, training job was a section inside the rating job which depending to the browser cookies of worker (i.e. cookie expired means training is required) was visible or hidden. After each time the training section was shown, the cookie expiry time was renewed.

Rating job: 10 stimuli (and one trapping stimulus) were assessed in one rating task. In case that training was necessary, five pre-selected stimuli were shown in the corresponding section. They cover entire range of scale. The training certificate (cookie) expires after one day. For the environmental test, four pairs of stimuli were previously selected. Each pair presents a difference of 0.6MOS in laboratory test. In each question, the worker should select the stimulus with better quality from the presented pair (or difference is not detectable) as suggested in Appendix III of draft text for P.CROWD recommendation. The threshold of 0.6MOS was selected as a result of laboratory experiment using adaptive psychophysical method (staircase) when normal hearing subjects were participated in a study conducted in a typical calm room condition and using a common listening device. In addition, a math question was used to check if workers use both earplugs. Workers were also forced to listen to each stimulus until the end before casting their vote. Workers were able to listen to each stimulus as many time as they wish but the number was captured. The dataset contains 1152 (48 x 24) files which was divided into 116 stimuli sets (each contain 10 randomly selected files). It was planned to collect 10 votes per set.

Data screening: In total 1160 response packages (each include 10 votes) from 71 unique workers were collected. From them, 12 response packages with a wrong answer either to the math question or to the trapping questions were removed. In addition, 106 other response packages were removed as workers failed in the environmental screening test. Overall, 10420 votes were accepted from 68 workers and used for further analyses.

CS 501

For this crowdsourcing test, we used clickworker[3], a German based crowdsourcing platform in which their workers are mainly from Germany, Austria and Belgium. Clickworker does not support audio playback (as of April 2018), thus a HTML JavaScript based framework[4] was implemented to administer the test to the workers with a Node.js server for the data collection.

[3] https://www.clickworker.com

[4] https://gitlab.com/zequeira/SQAT-Cr.git

The crowdsourcing study was designed with the P.CROWD draft recommendation in mind. It contained three phases, i.e. Qualification, Training and Rating. In the following we outline the differences in each of the phases to those defined in P.CROWD.

Qualification phase: The Qualification phase in our crowdsourcing study included three passages in German to evaluate the workers' German knowledge. Once they heard each audio, they were asked to select the right statement (related to the passage heard) out of three that were provided.

Training phase: The Training phase permitted to control the use of headphone (in-ear or over-ear), and a short math exercise including digits panning left to right in stereo controlled for its two-eared use. In addition, workers were presented with a short hearing impairment test. 10 different audio files were created containing white noise at different frequencies ranges. Thus, listeners were presented with four octave filters around the frequencies: 500Hz, 1Khz, 4Khz and 8Khz, at a -46dBov,-66dBov and -76dBov (just for 1Khz and below) level. Workers were then asked whether they heard something in each audio file. Furthermore, listeners listened to five speech stimuli (taken from the dataset) that covered the entire MOS range, so they could get to know what to expect on the Rating part and also get familiar with the scale. When answering correctly the math trapping question, workers were assigned a qualification for one-hour time frame in which they could perform the rating job at which.

Rating phase: The rating phase included 15 stimuli. Work in [6] points out that is desirable to offer tasks with less speech stimuli in order to increase the listener retention and decrease the study turnaround time. In addition, one trapping question (created as P.CROWD recommended) was inserted randomly within the first five stimuli and one between the 10th and the 15th speech sample. After having listened to all stimuli, workers were asked to state in a slider how much fatigued they felt. When listeners failed any of the trapping questions, the access to the rating job was revoke. In addition, environmental background noise was recorded when workers played the first and the ninth sample (7.5 seconds each time). Workers were not able to provide their opinion on the scale unless they listened first to the speech sample. They could not go forward until the audio was played completely and an option selected on the scale. And they could listen to each speech sample as many times as they wished.

Data screening: We collected a total of 5245 ratings from 64 unique listeners. All of them answered properly the included trapping question. No crowdworker was removed because of hearing test as: 1) there was no guarantee that workers reported "hearing" a noise, they really heard it, 2) hearing/not hearing the noise could be due to the listening level of device under used. In addition, 136 ratings were identified as extreme outliers (beyond an outer fence of boxplot) and removed.

Table 2 summarize the abovementioned crowdsourcing tests.

Table 2: Summary of conducted crowdsourcing tests

Study name	**CS401**	**CS501**
Original dataset	401	501
Crowdsourcing test		
Experimenter group	QUL1	QUL2
Date	March 2018	February 2018

Crowdsourcing platform	MTurk	Clickworker
External framework	No	Yes
Duration	3 days	11 days
Number of crowd workers[5]	68	64
Votes per File (M / SD)	9 / 1.2	25.5 / 3.5
Votes per Condition (M / SD)	217 / 4.8	102.2 / 7.3
Votes by CS worker (M / SD)	146.6 / 139	79.8 / 48.5
Files rated in one session	10	15
Method of CS test		
Workers were check for..		
being a native	Filtering by location	German Test
listening impairment	Asking and digit triple test (threshold -11.2dB SNR)	Web hearing test (white noise at different dB)
environment	Four pair comparison at the beginning of session	Environment Background noise recording
Further validity check method		
Attention questions per task	1	2
Sessions removed (outliers)	118	136

Results

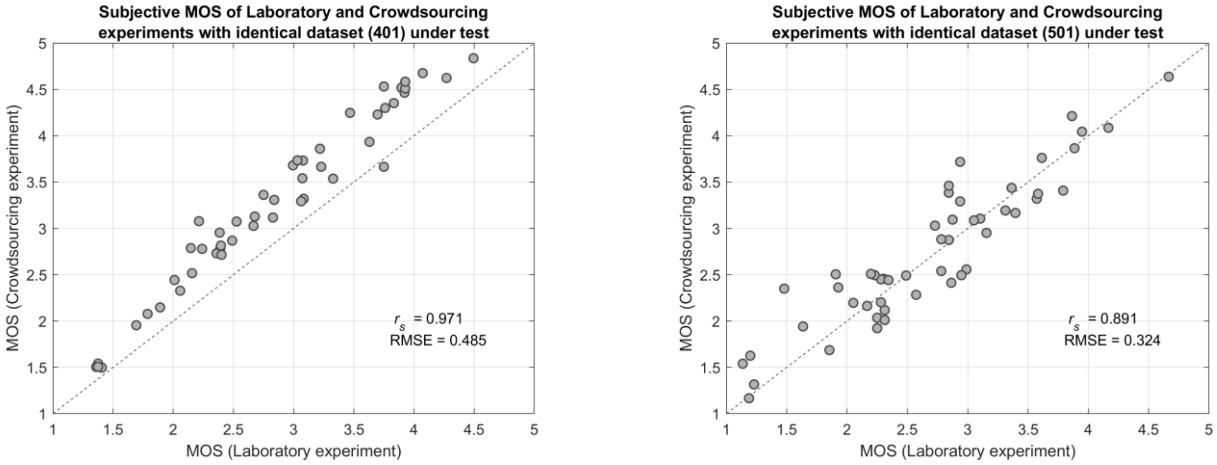
For each crowdsourcing test, subjective mean opinion scores (MOS), standard deviations and 95% confidence intervals were calculated per stimulus and per condition. The MOS values per condition obtained from the crowdsourcing tests were compared with the values provided from the corresponding laboratory based experiments (see Table 3).

Table 3: Comparison between MOS values obtained in CROWDSOURCING study with MOS values reported by laboratory study using a same dataset

Test	r_s	P	RMSE	Overlapping 95%CI (CS-Lab)	Votes per condition M (STD)
CS401	.971	<0.001	0.485	4 / 48	217 (4.8)
CS501	.891	<0.001	0.324	35 / 50	102 (7.3)

Results show that there is a high correlation between MOS values obtain through crowdsourcing test and those provided in the laboratory test. Figure 1-a shows that for study CS401, there is a bias and different gradient between crowdsourcing and laboratory scores.

[5] Workers with one or more accepted responses to the rating job.



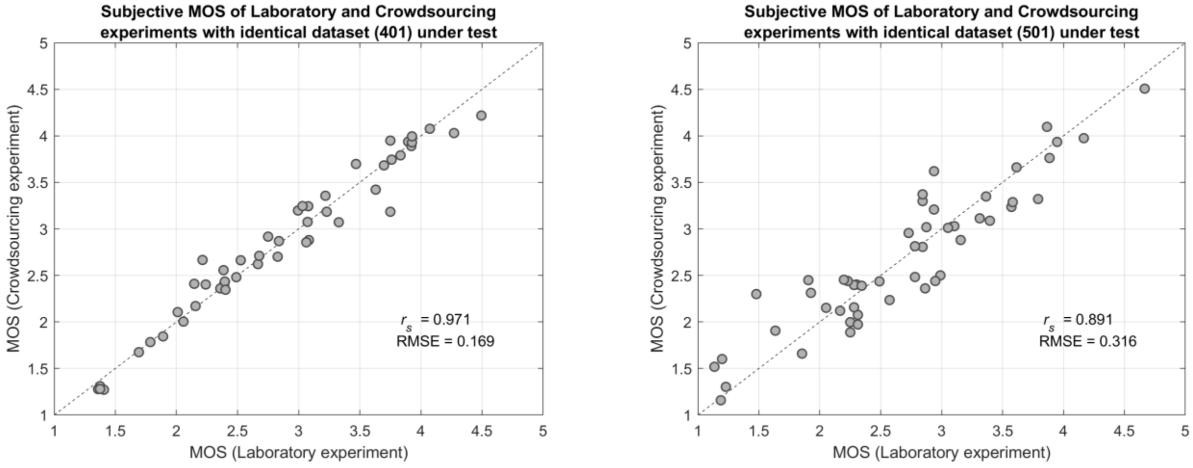
(a) (b)

Figure 1: Comparison between MOS values calculated per degradation conditions. (a) CS401, (b) CS501.

Applying a first-order mapping significantly decreases the RMSE and increase the number of conditions with overlapping 95%CI for CS401 (Table 4, and Figure 2).

Table 4: Comparison between MOS values obtained through CROWDSOURCING approach and laboratory after applying 1st order mapping

Test	r_s	P	RMSE	Overlapping 95%CI (CS-Lab)	1st order mapping b0,b1
CS401	.971	<0.001	0.169	37 / 48	-0.05, 0.882
CS501	.891	<0.001	0.316	34 / 50	0.032, 0.965



(a) (b)

Figure 2: Comparison between MOS values after applying 1st order mapping.

Discussion

Results show that there is a high correlation between the MOS values obtained in the laboratory conditions and MOS values obtain through the proposed approach by P.CROWD when a same dataset are used. In study CS401, systematic offset and gradient difference between laboratory and crowdsourcing MOS values were observed. Based on [5] the offset can be due to "overall" quality that is presented during the experiment and/or different listening gear or environmental noises. We expect that the observed offset in study CS401 was due to the fact that all workers did not rate all entire range of available stimuli. We expected that this effect should be vanished by appropriated training session however the re-training sessions was forced after 24hours (practically just 2 workers trained more than once). The different in gradient could also have a same origin: as a worker did not rate samples that cover the entire range, he/she may tend to use the entire scale for the range of quality that is included. We recommend to use a shorter temporal qualification of 1 hour as mentioned in the P.CROWD. The observed offset and gradient were corrected by 1st order mapping.

In addition, we recommend further investigation on practical methods to test suitability of environment and self-screen hearing test.

Bibliography

[1] Buschermöhle, M., Wagener, K., Berg, D., Meis, M. and Kollmeier, B. 2015. The German Digit Triplets Test (Part II): Validation and Pass/Fail Criteria. 54, 1 (2015), 6–13.

[2] Draft Recommendation ITU-T P.CROWD Subjective evaluation of speech quality with a crowdsourcing approach.

[3] Naderi, B. 2018. *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer International Publishing.

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

[5] Recommendation ITU-T P.1401 (2012). Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.

[6] Zequeira Jiménez, R., Fernández Gallardo, L. and Möller, S. 2018. Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing. *accepted for: 10th International Conference on Quality of Multimedia Experience (QoMEX)* (2018).

______