Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing. A Speech Quality Assessment Case Study

Rafael Zequeira Jiménez
Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany
rafael.zequeira@tu-berlin.de

Laura Fernández Gallardo
Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany
laura.fernandezgallardo@tu-berlin.de

Sebastian Möller
Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany
sebastian.moeller@tu-berlin.de

ABSTRACT

Crowdsourcing provides an exceptional opportunity for the rapid collection of human input for data acquisition and labelling. This approach have been adopted in multiple domains and researchers are now able to reach a demographically diverse audience at low cost. However, it remains the question of whether the results are still valid and reliable. Previous work have introduced different mechanisms to ensure data reliability in crowdsourcing. This work examines to which extend, “trapping question" or “outliers detection" assure reliable results to the detriment of, overloading task content with stimuli that are not of interest for the researcher, or by discarding data points that might be the true opinion of a worker. To this end, a speech quality assessment study have been conducted in a web crowdsourcing platform, following the ITU-T Rec. P.800. Workers assessed the speech stimuli of the database 501 from the ITU-T Rec. P.863. We examine results' validity in terms of correlations to previous ratings collected in laboratory. Our outcomes shows that neither of the techniques under investigation improve results accuracy by itself, but a combination of both. Our goal is to provide empirical guidance for designing experiments in crowdsourcing while ensuring data reliability.

CCS CONCEPTS

• **Human-centered computing → User studies**; *Field studies*; *Web-based interaction*; *Empirical studies in HCI*; • **General and reference** → Reliability;

KEYWORDS

outliers detection; trapping questions; crowdsourcing; speech quality assessment; gold-standard questions; users reliability; data validity

ACM Reference Format:
Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. 2018. Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing. A Speech Quality Assessment Case Study. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191545

1 INTRODUCTION

The crowdsourcing (CS) paradigm offers small tasks to anonymous users on the Internet that normally require human intelligence for being resolved. The users (crowd-workers) can perform such tasks from their computer or mobile device and get rewarded after completion. Nowadays, CS has been adopted in multiple domains and researchers have found a fast, low cost, and scalable method to gather more labeled data than traditional approaches.

Subjective speech quality assessment experiments have been traditionally conducted under controlled laboratory conditions with high-end equipment. This approach permits an optimal control over the study setup to the detriment of the number of participants. In contrast, CS provides the means to reach a wider and diverse audience to collect quality ratings at a fraction of the cost and time than traditional practices of in-Lab annotations. However, it remains the question of whether the collected ratings in an online platform are still valid and reliable. CS has been typically found to deliver noisier data, therefore, different quality control mechanisms has been proposed to ensure reliable results and overcome this challenge [1, 4, 11]. Accurate speech quality ratings are of a main importance for telecommunication systems' providers, as such data is used to train models to predict the perceived quality of the different codecs applied to the speech signal. Thus, reliable ratings are essential for a proper evaluation of their systems.

The rest of the paper is structured as follows: the next section reviews existing work that used CS to assess the quality of different multimedia contents, and the impact of the quality control mechanism employed. Section 3 presents the experiment setup as well as the database employed. The results of contrasting the laboratory (Lab) with the CS outcomes are exposed in Section 4 with our approach to ensure accurate results. Finally, Section 5 concludes and outlines our directions for future work.

2 RELATED WORK

Trapping questions (TQ) can be seen as gold standard questions that can be used to identify inattentive or willfully cheating workers. Work in [9] used mobile-CS to collect quality ratings of speech samples. The authors evaluated different types of TQ and examined their influence on the gathered ratings. An increase in the correlation was found between the MOS ratings collected in the Lab and

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191545

the CS results when employing TQ ($\rho = 0.886$ to $\rho = 0.909$). However, the authors did not apply any outliers detection mechanism, and quality control relied on just discarding the ratings from the workers deemed unreliable by the TQ setup.

[2] investigates the viability of a web-CS platform for the subjective evaluation of audio with intermediate impairments. The authors used a screening task as control mechanism to account for workers' hearing abilities and listening environments. Results in terms of overall audio quality were correlated ($\rho = 0.78$) to previous ratings collected in lab.

Research in [5] proposes a CS test methodology to assess the user perceived quality of Internet applications like YouTube. The authors employed gold standard questions to identify unreliable workers. Their approach lead them to improve significantly the intra- and inter- rater reliability by discarding the ratings from untrustworthy workers. Yet, such a filtering technique reduced the number of valid crowd-workers by approx. 25% and made them discard three fourth of the subjective ratings collected. As well authors of [10] reduced the number of workers by 34.3% to be considered for a CS study in image aesthetic appeal. They filtered out the outliers following a multi-fold technique, and removed users based on verification questions.

3 EXPERIMENT SETTINGS

This paper investigates whether outliers detection, trapping questions or a combination of both, improve the results accuracy in the context of a speech quality assessment task. We also determine whether it is possible to optimize such quality control mechanism without discarding a big number of users and/or ratings, which may lead to poor performance in terms of costs and time.

The analysis is conducted on data from our prior work in [14], which analyses the influence of the number of presented speech stimuli on the reliability of listeners' ratings. Thus, a CS study was conducted with 53 workers. They were asked to rate speech stimuli with respect to their overall quality on a 5-point scale in accordance with the ITU-T Rec. P.800 [6]. We examine results' accuracy in terms of correlations to previous ratings collected in Lab.

3.1 Speech Material

The stimuli employed in our study were taken from the database number 501 from the ITU-T Rec. P.863 [7] competition. This database contains different degradation conditions in accordance to the ITU-T Rec. P.863. Four native German speakers were recorded per condition uttering four different sentences in German. In total, 200 speech files (8s to 10s long) were arranged accounting for 50 degradation conditions, e.g. different audio bandwidths, temporal clipping, speech coding at various bitrates, diverse types of ambient background noise, frequency distortions and, combinations of these degradations.

The database contains subjective quality assessments to the 200 stimuli made by 24 different native German listeners, in accordance with the ITU-T Rec. P.800 [6]. The Mean Opinion Scores (MOS) for each stimulus are taken as a reference for the analysis presented in this paper (from now on referred as "Lab-MOS"). A Kendall's ζ coefficient [8] calculated among the listeners, revealed a statistically significant agreement between the Lab participants when assessing all the speech stimuli $W = 0.614$ ($p < 0.001$). This coefficient can only be calculated if all samples are presented to all listeners, which is not the case in CS.

3.2 Crowdsourcing Study

We used the clickworker[1] CS platform to conduct our experiment. Most of its users are from German speaking countries, which was a good fit for our study needs. However, clickworker does not support for audio playback (as of February 2018), we then created a HTML JavaScript based framework to administer the test to the workers. We used crowdcrafting [2] as interface to display the test and a Node.js server for the data collection.

Inspired by work in [9, 13], our CS experiment included a Qualification phase that permitted to adjust the device volume to a comfortable level, and regulated the use of headphones. A short math exercise with digits panning left to right in stereo controlled for two-eared usage. Six different audio files were prepared for this and presented randomly every time the worker executed the Qualification. This questions served as well as quality control to prevent inattentive workers from participating in our study during 12 hours.

Upon successful execution of the Qualification phase, or after 12 hours in case of failure, workers were presented immediately with the speech quality assessment task (SQAT). The study architecture can be seen in Figure 1. The SQAT permitted the listeners to assess the overall quality of 20 speech samples on a 5-point scale, see Figure 2. Workers could not provide their opinion on the scale unless they listened first to the speech sample. They were not able to go forward until the audio was played completely, and they could listen to each speech sample as many times as they wished.

To evaluate the entire dataset, the listeners could participate in the study up to 10 times (with the restriction of only one execution every 12 hours). In addition, two trapping questions (TQ) were inserted randomly within the first five stimuli from every ten speech samples. Like in [9], these TQ presented a stimulus taken from the database but interrupted after four seconds, listeners were then informed about the importance of their work and were asked to select an specific item on the scale. The TQs' GUI was the same as the rest of presented stimuli, see Figure 2. When workers failed to answer correctly the TQ, the ratings from the set of those ten stimuli where considered unreliable and discarded. More details on the Qualification and the SQAT can be found in [14].

4 RESULTS

87 workers in total participated in the CS study. 8 of them answered wrong the Qualification phase and were granted with a 12 hours window that prevented them from conducting our experiment. From those 8 workers, 6 returned back to our study and provided reliable ratings. Overall, 53 crowd-workers (balanced age) yielded 4840 ratings, more details on their demographics is presented in Table 1. Surprisingly, all the workers answered correctly the TQ in the SQAT, and all the 4840 ratings were deemed reliable. The rest of the listeners either started out the study and did not finish,

[1] https://www.clickworker.com last accessed February 2018

[2] https://crowdcrafting.org last accessed February 2018

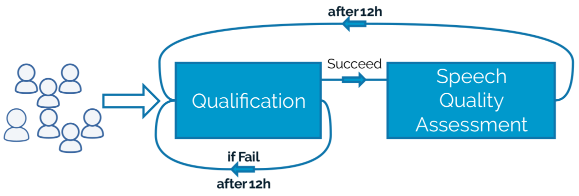
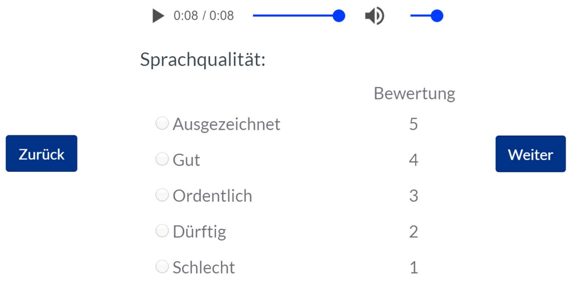


Figure 1: Study architecture, workers could participate only once every 12 hours.


▶ 0:08 / 0:08

Sprachqualität:

	Bewertung
○ Ausgezeichnet	5
○ Gut	4
○ Ordentlich	3
○ Dürftig	2
○ Schlecht	1

Figure 2: Graphical interface presented to the workers for the SQAT. The text translate from German: “Speech Quality" and “Rating". The scale (in descending order): “Excellent", “Good", “Fair", “Poor" and “Bad".

Table 1: Demographic information of the 53 workers that executed properly the SQAT. Values are expressed in percentages. “NP" stands for *“Not Provided"*, e.g. workers did not provide that information.

Language		Country		Gender	
German	96.2	Germany	98.1	Male	60.4
NP	3.8	Austria	1.9	Female	39.6

or dropped off after reading the instructions (we don't have an explanation for this behavior).

The 4840 gathered ratings account for 24 to 26 assessments made by different listeners on each of the 200 speech stimuli. To determine the validity of this gathered data, a Spearman's rank-order correlation was run to assess the relationship between the Laboratory and the CS ratings. Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of a scatter-plot. Also, the Root Mean Square Error (RMSE) was calculated between the ratings in the Lab and in CS. Strong positive correlation was found between the Lab-MOS and the CS-MOS, $\rho = 0.864$ ($p < .001$), as well as a low $RMSE = 0.474$. This outcome motivates the use of CS for collecting reliable annotations of speech quality, as an alternative to a more controlled environment like in Lab test.

Next, we examine whether filtering out the contributions from “unreliable workers" might improve or not the accuracy of our results. Work in [4, 9, 13] recommends the use of TQ or other control mechanisms as a mean to identify untrustworthy crowd-workers and discard all of their answers. Such a technique resulted to be effective in [9] and improved slightly the results in [13].

In this work, we labeled a worker as unreliable or untrustworthy when s/he failed the TQ in the SQAT, or the Qualification phase more than once. Figure 3 exposes such cases (workers are assigned with *W1* to *W8*). Thus, when discarding the contributions from *W4*, *W5* and *W7* (320 ratings in total) (*W6* did not executed the SQAT), we account for 4520 ratings in total, which represent 21 to 25 assessments per speech sample made by different listeners. We call this method: *“filtering by trapping question"* (F-TQ). Calculating the Spearman's correlation under these conditions throw a slightly decreased score of: $\rho = 0.862$ ($p < .001$), and even worst if we discard all the workers that failed the Qualification (F-TQ'): $\rho = 0.854$ ($p < .001$).

Filtering our data based on the TQ from the Qualification phase did not improved the results. This outcome shows that this method is not always valid, and workers failing once a quality control mechanism might actually provide reliable ratings in further executions of the study.

Moreover, we investigate whether applying outliers detection to our data improves the results. According to the labeling rule introduced in [3], we removed the ratings with a distance from the median higher than $2.2 \cdot IQR$ (interquartile range). We executed this analysis for each speech stimuli of the dataset and 122 ratings were identified as extreme outliers and were discarded. We refer to this method as: *“filtering by outliers detection 1"* (F-OD1). Box-plots in Figure 4 represents ten speech stimuli chosen arbitrarily to showcase some of the outliers and extreme outliers in our study. The resulting Spearman's correlation after applying F-OD1 was then: $\rho = 0.863$ ($p < .001$), still not better than the first coefficient calculated when no data was removed.

In addition, we eliminated all the ratings (1480 in total) from the 12 workers that were identified as outliers according to [12] (e.g. their ratings were outliers three times or more). We call this method *“filtering by outliers detection 2"* (F-OD2). An improve in the correlation coefficient was seen this time: $\rho = 0.867$ ($p < .001$).

Finally, we examine whether combining F-TQ, F-OD1 and F-OD2 leads to even more accurate results. We refer to this approach as: F-TQ-OD. To this end, we applied F-OD1 and F-OD2 to our data and discarded 1529 ratings. In addition, we identified the outliers made by all workers that failed the TQ in the Qualification phase. This analysis led us to drop 17 data points more. The resulting 3294 ratings represent 13 to 21 assessments from different listeners to each of the 200 speech files. The Spearman's rank-order correlation to the Lab-MOS is now higher compared to those previously calculated: $\rho = 0.868$ ($p < .001$). Table 2 presents a summary of the correlations achieved with each of the methods, and the ratings discarded in each case from the amount of 4840 initially collected.

F-TQ-OD slightly outperforms the approach in [5] and in [10] in terms of the amount of workers and ratings discarded. Applying F-TQ-OD in our study lead to filter out 22% of the crowd-workers and 31.9% of the collected ratings: less than in [5], where 25% of the workers and 75% of the ratings were discarded, and also less in comparison to [10], where 34.3% of the workers were removed from the final analysis.

While F-TQ-OD led to more accurate results, it was at the expenses of discarding a considerable amount of data points. This translates in an increase of the experiment cost in case a certain number of assessments on the dataset is needed. However, this

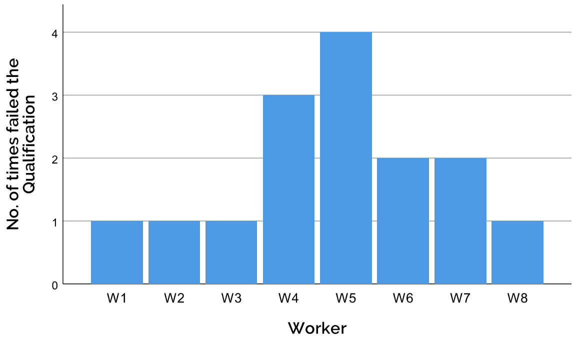


Figure 3: Amount of times a worker failed the Qualification phase.

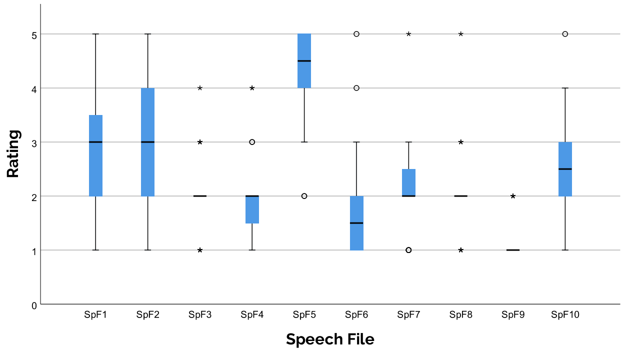


Figure 4: Boxplots of ratings for ten speech stimuli of the dataset. Lines extend to 1.5 times the interquartile range, circles indicate outliers, asterisks illustrate extreme outliers and notches depict 95% confidence intervals.

Table 2: Correlation (ρ) and Root Mean Square Error (RMSE) between the Lab-MOS and the CS-MOS when filtering by outliers according to [3] and [12] (F-OD1 and F-OD2, respectively), and when filtering by TQ (F-TQ).

Method	Ratings discarded	ρ	*RMSE*
-	0	0.864*	0.474
F-TQ	320	0.862*	0.476
F-TQ'	780	0.854*	0.480
F-OD1	122	0.863*	0.477
F-OD2	1480	0.867*	0.474
F-TQ-OD	1546	0.868*	0.479

$^*p < 0.001$

outcome also suggests that valid results are also possible with less iterations on the data (between 13 and 21 in our study).

5 CONCLUSIONS

This work, proposes a method to filter unreliable data gathered in Crowdsourcing studies. Our method combines outliers detection mechanisms [3, 12] with trapping questions to identify invalid data points and untrustworthy workers. Our approach have been tested with 4840 ratings collected in a speech quality assessment study conducted in a crowdsourcing platform. While the proposed method outperforms outliers detection and trapping questions when applied independently, further testing would be required to understand to which extend this mechanism can be applied, and for which types of experiments.

REFERENCES

[1] Daniel Archambault, Helen C Purchase, and Tobias Hoßfeld. 2017. *Evaluation in the Crowd: An Introduction.* Springer International Publishing, Cham, 1–5.

[2] Mark Cartwright, Bryan Pardo, Gautham J. Mysore, and Matt Hoffman. 2016. Fast and Easy Crowdsourced Perceptual Audio Evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 619–623.

[3] David C Hoaglin and Boris Iglewicz. 1987. Fine-tuning some resistant rules for outlier labeling. *J. Amer. Statist. Assoc.* 82, 400 (1987), 1147–1149.

[4] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel. 2014. Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing". (oct 2014).

[5] Tobias Hoßfeld, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. 2011. Quantification of YouTube QoE via Crowdsourcing. In *2011 IEEE International Symposium on Multimedia.* 494–499.

[6] ITU-T Recommandation P.800. 1996. *Methods for subjective determination of transmission quality.* International Telecommunication Union, Geneva.

[7] ITU-T Recommandation P.863. 2014. *Perceptual objective listening quality assessment.* International Telecommunication Union, Geneva.

[8] Maurice George Kendall. 1970. *Rank Correlation Methods* (4th ed.). Charles Griffin.

[9] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, and Sebastian Möller. 2015. Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm. In *Interspeech.* ISCA, 2799–2803.

[10] Judith Redi and Isabel Povoa. 2014. Crowdsourcing for Rating Image Aesthetic Appeal: Better a Paid or a Volunteer Crowd?. In *Proceedings of the 2014 International {ACM} Workshop on Crowdsourcing for Multimedia (CrowdMM '14).* 25–30.

[11] Judith Redi, Ernestasia Siahaan, Pavel Korshunov, Julian Habigt, and Tobias Hoßfeld. 2015. When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal. *Fourth International Workshop on Crowdsourcing for Multimedia* (2015), 33–38.

[12] Barbara G Tabachnick and Linda S Fidell. 2007. *Using multivariate statistics.* Allyn & Bacon, Pearson Education.

[13] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. 2017. Scoring Voice Likability using Pair-Comparison: Laboratory vs. Crowdsourcing Approach. In *Ninth International Conference on Quality of Multimedia Experience (QoMEX).* 1–3.

[14] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. 2018. Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing. In *accepted for: 10th International Conference on Quality of Multimedia Experience (QoMEX).*