Environmental Noise Recording as a Quality Control for Crowdsourcing Speech Quality Assessments

Rafael Zequeira Jiménez, Laura Fernández Gallardo, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin, Germany
Email: rafael.zequeira@tu-berlin.de, laura.fernandezgallardo@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

The Crowdsourcing (CS) paradigm offers small tasks to anonymous users on the Internet. Human-centered speech quality assessment studies have been traditionally conducted under controlled laboratory conditions. Nowadays, CS provides an exceptional opportunity to transfer such experiments to the internet and reach a wider and diverse audience. However, data from CS can be corrupted due to users' neglect and hence quality control mechanisms are required to ensure reliable outcomes. While previous works have presented trapping questions or majority voting to ensure good results, this work introduces user-environmental noise recording to discard unreliable users located in noisy places. To this end, a speech quality assessment study was conducted in the clickworker CS platform. The speech stimuli are taken from the database 501 from the ITU-T Rec. P.863 and the results are to be contrasted to the existing lab ratings. This work analyzes whether environmental noise recording can be used to identify unreliable workers. Furthermore, the effects of discarding users deemed untrustworthy on the correlation between the CS and the Lab results is studied. Our outcomes highlight the importance of controlling for users' background noises to ensure reliable results in speech quality assessments conducted via CS.

Introduction

Crowdsourcing (CS) has emerged as a competitive tool to conduct user studies on the Internet. The users in CS (also called workers or crowd-workers), execute small tasks remotely from their computer or mobile device in exchange for a monetary compensation. This approach has been adopted in multiple domains as a mean to collect human input for data acquisition and annotations. Researchers are now able to address their user tests to a wider and diverse audience while reducing turnaround time and costs. However, it remains the question of whether the collected ratings in an online platform are still valid and reliable, that is, comparable to those gathered in a constrained Laboratory (Lab) environment. There is a lack of control to supervise the participant, and often not enough information on their playback system and background environment. Therefore, different quality control mechanisms has been proposed to ensure reliable results and to monitor these factors to the extent possible [1, 2].

The quality of transmitted speech signals is of main importance for telecommunication network providers, as it is one of the main indicators to evaluate their systems and services. Traditionally, subjective speech quality studies are normally executed under controlled conditions with professional audio equipment. This way a good control over the experiment setup can be achieved but with some mayor disadvantages, e.g. it is expensive, time consuming, and often the number of participants is rather low. Consequently, the outcomes might not be representative enough for all users. In contrast, CS represent a promising approach for the rapid collection of quality ratings at a fraction of the cost and time.

Nevertheless, the implementation of existing subjective testing methodologies into an Internet-based environment is not straightforward. Multiple challenges arise that need to be addressed in order to gather valid results, e.g. task length and presentation, user motivation and reliability. This work investigates weather environmental noise recording can be used as a mechanism to identify unreliable workers in CS. To this end, a speech quality assessment study have been conducted in a web CS platform, following the ITU-T Rec. P.800 [3]. 64 workers were asked to rate speech stimuli with respect to their overall quality on a 5-point scale. Environment background noise was recorded while listeners executed the test, and the collected files were analyzed to identify wrong executions of the test. We validate our results in terms of correlations to previous ratings collected in Lab.

The rest of the paper is structured as follows: the next section reviews existing work that used CS for speech related tasks, and outline the quality mechanism their employed. Next, we presents the experiment setup as well as the database employed. The results of contrasting the laboratory (Lab) with the CS outcomes are exposed in the "Results" section. Finally, in the "Conclusion", we outlines our directions for future work.

Related Work

CS has been typically found to deliver noisier data, thus, multiple techniques have been proposed in the literature to achieve reliable results. Authors in [4] used mobile-CS to determine the influence of different trapping question (TQ) on the ratings in a speech quality assessment task. A strong correlation ($\rho = 0.909$) was found between the MOS ratings collected in the Lab and the CS results for the TQ in which a recorded voice was presented in the middle of a random stimuli, and motivated the worker to stay focus and produce quality work.

Research in [5] employed web-CS to collect ratings of users' perception of naturalness of synthesized speech in a discrete 5-point scale. Authors pointed out that workers in CS have little incentive to submit intentionally inconsistent result, since they might get blocked by requesters and banned from participating in any future study. Still, the authors used a screening process to detect and discard inaccurate or malicious submissions.

[6] analyses how people value speech-based take-over requests as a function of speech rate, background noise and speakers' gender and emotional tone. Workers in CS listened to ten speech recordings and they were asked to give scores about the urgency, pleasantness and commanding, characteristics of every sample. In addition, a control question about the speakers' gender was included to catch inattentive workers. Authors reports on the effect of noise level on whether the message in the speech was easy to understand for the crowd-workers.

Method

Speech Material

The stimuli employed in our study were taken from the database number 501 from the ITU-T Rec. P.863 [7] competition, which contains multiple degradation conditions. Four native German speakers were recorded per condition uttering four different sentences in German. In total, 200 speech files (9s on avg.) were arranged accounting for 50 degradation conditions, e.g. temporal clipping, different audio bandwidths (narrowband, wideband, super wideband), frequency distortions, speech coding at various bitrates, diverse types of ambient background noise, 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 [3]. 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”).

Crowdsourcing Study

For our study we used the clickworker[1] CS platform. Their users are mainly from Germany, Austria and Belgium, thus a good fit for our experiment needs. In addition, we implemented a HTML JavaScript based framework to carry out the test and a Node.js server for the data collection.

Based on work in [4, 8], our CS experiment included a Qualification phase that permitted to adjust the workers' device volume to a comfortable level while listening to an audio file. They were also presented with a short math exercise with digits panning left to right in stereo, this way we controlled for the headphones' two-eared usage. When workers failed to answer correctly this control question, they were prevented from conducting the study during 12 hours, otherwise they were redirected immediately to the speech quality assessment task (SQAT). The study architecture can be seen in Figure 1.

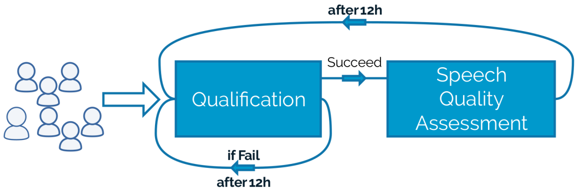


Figure 1: Study architecture, workers could participate only once every 12 hours.

The SQAT permitted the listeners to assess the overall quality of 40 speech samples on a 5-point scale, see Figure 2. In addition, we remember them to grant access to their microphone when requested so we could record their environment background up to 15 seconds for quality control. 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.

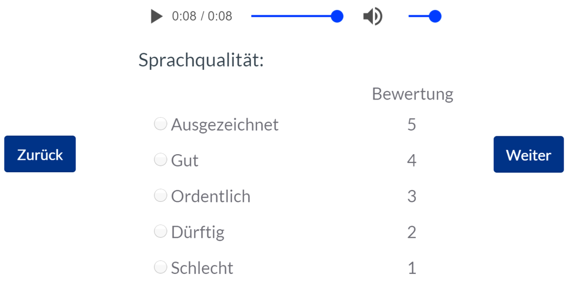
Workers could participate in our study up to 5 times to evaluate the entire dataset (with the restriction of only one execution every 12 hours). . We inserted one TQ randomly within the first five stimuli from every ten speech samples to catch sloppy listeners. These TQ presented a stimulus similar to the rest of stimuli but interrupted after four seconds, workers were then informed about how much we value their work and were asked to select an specific item on the scale [4]. The TQs' GUI was the same as the rest of presented stimuli, see Figure 2. More details on the Qualification and the SQAT can be found in [9].

Results

The Qualification was effective and prevented 9 workers from participating in our study during 12 hours. 5 of them returned back and provided reliable ratings. In total, 64 crowd-workers (balanced age) produced 5080 ratings which accounts for 25.4 assessments on average (range: 23–29) made by different listeners on each of the 200 speech stimuli. Details on their demographics are presented in Table 1.

We executed an outliers detection analysis according to the labeling rule proposed in [10], and removed 67 ratings identified as extreme outliers, that is, ratings located at a distance from the median higher than $2.2 \cdot IQR$ (interquartile range). A Spearman's rank-order correlation was run to assess the relationship between the Laboratory and the CS ratings. A visual inspection of a scatter-plot showed the relationship to be monotonic. We also calculated the Root Mean Square Error (RMSE) between the ratings in the Lab and in CS. Strong positive and significant correlation was found between the Lab-MOS and the CS-MOS, $\rho = 0.888$ ($p < .001$), as well as a low $RMSE = 0.409$.

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



Sprachqualität:

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

Figure 2: Graphical interface presented to the workers for the SQAT (in line with [3]). The text translate from German: "Speech Quality" and "Rating". The scale (in descending order): "Excellent", "Good", "Fair", "Poor" and "Bad".

Table 1: Demographic details of the 64 workers that executed properly the SQAT. Values are expressed in percentages.

Language		Country		Gender	
German	98.4	Germany	92.2	Male	57.8
Turkish	1.6	Austria	7.8	Female	42.2

Environmental Recordings

A JavaScript code within the SQAT permitted us to record remotely the environment background noise of the workers while their conducted our study. 7.5 seconds were recorded when they listened the first and the ninth speech stimulus accounting for two files per test session. In total 185 environmental recordings were collected from 44 different listeners. We hypothesize that the rest of the workers either decided to do not grant access to their microphone or an error on their browser prevented the recording from happening.

The gathered files were analyzed to detect wrong executions of our study, e.g. workers conducting the test using loudspeakers instead of headphones or workers performing the test in noisy environments. We noticed that 18 listeners executed our experiment employing loudspeakers, all of the ratings (1200 in total) from each of those test sessions were labeled as "performed using loudspeakers". Moreover, 38.9% of those 18 workers were outliers (e.g. provided ratings located at a distance from the median higher than $2.0 \cdot IQR$), or extreme outliers. A cross match between the workers deemed extreme outliers and the ones employing loudspeakers, revealed 3 untrustworthy listeners and all of their ratings were removed (432 in total).

The Spearman's rank-order correlation and the RMSE was calculated with the resulting 4581 ratings to determine if our approach led us to more accurate results. A slight increase was achieve: $\rho = 0.891$ ($p < .001$) and $RMSE = 0.411$. This outcomes motivates the use of environmental noise recordings as a mechanism to detect unreliable workers in CS.

Differences in the CS Study Execution

Furthermore, we analyze whether there was a difference in the ratings between the group of workers that conducted the study employing loudspeakers (G31), and the group using headphones (G3). As previously pointed out, the speech material presented to the listeners contained 50 degradation conditions. For each of those, a Mann-Whitney U test [11] was run to determine if there were differences between the rating scores provided by the workers of G31 and G3. Distributions of the rating scores were similar for both groups, as assessed by visual inspection. The median (Mdn) of the rating scores were statistically significantly different in 11 conditions, using an exact sampling distribution for U [12]. Table 2 exposes these results (U, z and p values), the Mdn values for each of the 11 conditions of the Lab results are also included to be taken as a reference. Figure 3 presents the MOS scores of G31, G3 and the Lab (only 11 conditions) with 95% confidence intervals.

A closer look into the conditions revealed that, specially the speech stimuli with wideband (WB) and super wideband (SWB) characteristics were rated with a greater quality score by the workers that used headphones. As for the rest of conditions, the distortions it contained were so evident, that workers were able to discriminate between them regardless the audio equipment employed. Interestingly, the MOS of G31 for some of the conditions were closer to the Lab-MOS (e.g. condition number: 13, 20, 43, 44 and 49) than those of G3. Research in [5] pointed out that workers using loudspeakers present a smaller discrimination capacity and don't perceive certain characteristics of the speech signal.

Conclusion

This paper investigates whether environmental noise recording can be used as a mechanism to detect unreliable workers and improve results accuracy. To this end, a crowdsourcing study have been conducted with 64 crowd-workers that assessed speech stimuli with respect to their overall quality. Our results are highly correlated to previously collected laboratory ratings. Moreover, an analysis was presented to show if individual degradation types were perceived differently by workers conducting the study employing headphones compared to those using loudspeakers. Our results shows that listeners were able to discriminate among 78% of the degradation condition regardless the audio equipment employed. This outcome motivates the use of web-CS for speech quality assessment tasks.

References

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

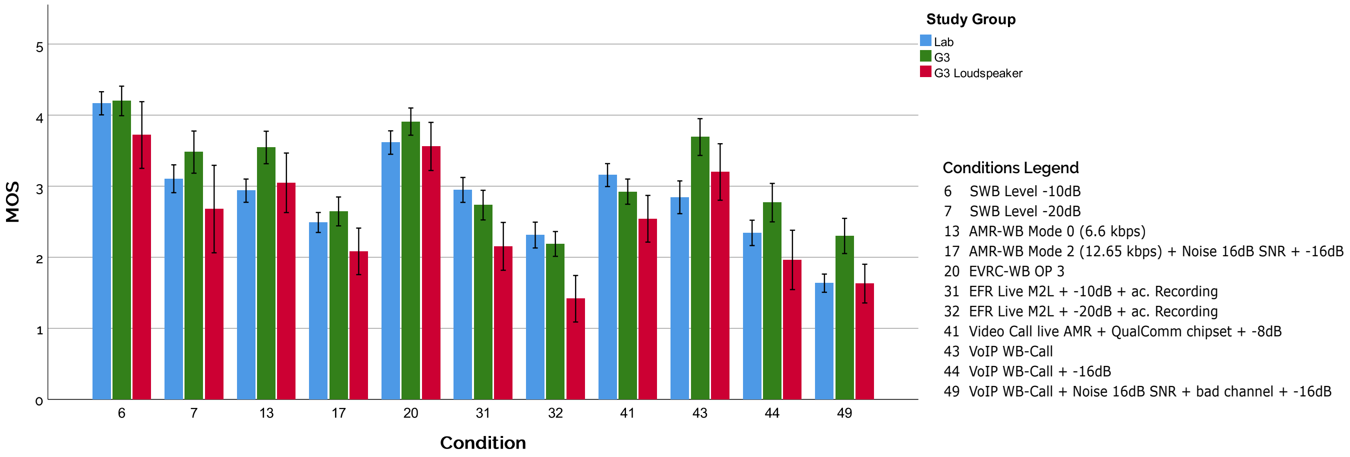


Figure 3: Comparison between the G31-MOS and G3-MOS with 95% confidence intervals. Represented as well the Lab-MOS to be taken as a reference. Information about the degradation conditions can be found in [7].

Table 2: 11 conditions for which the median (Mdn) of the rating scores were statistically significantly different between workers conducting the study employing loudspeakers (G31), and the ones using headphones (G3).

Cond.	U	z	p	Lab Mdn	G3 Mdn	G31 Mdn
6	752.0	-1.988	$= .142$	4	4	4
7	718.0	-2.355	$= .19$	3	4	2
13	551.0	-2.332	$= .02$	3	4	3
17	620.0	-2.509	$= .12$	2	3	2
20	725.0	-1.991	$= .47$	4	4	4
31	665.0	-2.821	$= .005$	3	3	2
32	484.0	-4.14	$< .001$	2	2	1
41	692.5	-2.048	$= .41$	3	3	2
43	761.0	-2.069	$= .039$	3	4	3
44	552.5	-3.215	$= .001$	2	3	2
49	599.5	-2.911	$= .004$	2	2	2

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

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

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

[5] Flavio P Ribeiro, Dinei A F Florêncio, Cha Zhang, and Michael L Seltzer, "CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, may 2011, pp. 2416–2419.

[6] P Bazilinskyy and J C F De Winter, "Analyzing crowdsourced ratings of speech-based take-over requests for automated driving," *Applied Ergonomics*, vol. 64, pp. 56–64, 2017.

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

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

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

[10] David C Hoaglin and Boris Iglewicz, "Fine-tuning some resistant rules for outlier labeling," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1147–1149, 1987.

[11] H B Mann and D R Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[12] L C Dinneen and B C Blakesley, "Algorithm AS 62: A Generator for the Sampling Distribution of the Mann- Whitney U Statistic," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 22, no. 2, pp. 269–273, 1973.