Effect of Environment in Speech Quality Assessment in Crowdsourcing

Rafael Zequeira Jiménez, Babak Naderi, Sebastian Möller

EFFECT OF ENVIRONMENT IN SPEECH QUALITY ASSESSMENT IN CROWDSOURCING

Rafael Zequeira Jiménez[◇] **Babak Naderi**[◇] **Sebastian Möller**[◇▷]

[◇] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[▷] DFKI Projektbüro Berlin, Berlin, Germany
rafael.zequeira | babak.naderi | sebastian.moeller @tu-berlin.de

ABSTRACT

The quality of the speech signal is of main importance as it influences the user experience of interactive systems such as: personal assistants, telephony calls, virtual conversational agents, and others. Speech quality studies have traditionally been conducted in constrained laboratory rooms with professional audio equipment. Nowadays, crowdsourcing (CS) represent a valid alternative for the rapid assessment of large speech databases. However, the question remains regarding the influence of the listeners' environmental background noise, and context. This paper compares two speech quality studies that were conducted in the laboratory following international standards, i.e ITU-T Rec. P.800 (Lab-traditional) and ITU-T Rec. P.808 (CS-simulated), respectively. During the test, listeners were exposed to background noise at different levels, and web audio recordings were also collected. We found that there was a statistically significant interaction between environment (Lab and CS) and level of noise, on the speech quality ratings provided by the listeners when assessing only one out of the 15 conditions that were under test.

1. INTRODUCTION

Crowdsourcing (CS) has emerged as a competitive tool for conducting user-centered studies. In a CS paradigm, the users (often referred to as workers or crowd-workers), accomplish small tasks remotely from their Internet-enabled computer or mobile device, and they get compensated for their participation. This approach has been adopted in a variety of domains to collect human annotations and for data acquisition. Researchers in the field of multimedia have benefited from the possibility of reaching a wider and more diverse audience for their user studies.

The quality of the speech signal is an important metric used by telecommunication network providers to estimate the performance of their systems and services. Subjective studies to determine the quality of speech stimuli, have been traditionally carried out in constrain laboratory (Lab) environments with professional audio equipment. In that way, a fair control over the experiment can be accomplish but with some disadvantages, Lab studies are time consuming and expensive. Additionally, it might not represent real life situations due to its artificial nature. In turn, CS stands as a more ecologically valid and promising approach for the rapid collection of speech quality scores at a fraction of the cost and time.

However, the implementation of existing subjective testing methodologies into an Internet-based environment like CS is not straightforward. Multiple challenges arise that need to be addressed in order to collect valid and reliable results. For instance, there is a lack of control to supervise the crowd-workers, and not enough information about their playback system and background environment. Oftentimes, workers do not follow the given instructions and might execute crowdsourcing tasks in noisy environments. Which could compromise the experiments' results, specially in audio related tasks.

The aim of this paper is to investigate the effect of environment in speech quality assessment tasks in Crowdsourcing. Specifically, the influence of environmental background noise in the collected speech quality scores, and whether there is an interaction effect between background noise and listening test method, i.e. CS-test vs Lab-test.

2. STUDY SETUP

A study was conducted in the laboratory in which two different groups of participants were recruited. They were asked to rate the quality of speech files under different environment background noise conditions.

The speech quality assessment test was divided into three sessions. First, listeners conducted a standard P.800 [1] test without background noise. Afterwards, participants in Group A (GA) conducted the remaining two session also following the Recommendation P.800 but, under the influence of street background noise at two different levels. Contrary, listeners in Group B (GB) executed the remaining two sessions in accordance to the Recommendation P.808 [2] and also under the influence of two street background noise conditions. The speech stimuli were the same in each of the three sessions and the order of the last two sessions was randomized. Table 1 summarizes this information and presents the levels at which the background noise was played during each test session.

The background noise levels were measured with a dummy head from "HEAD acoustics GmbH". For the presentation and simulation of the noise we employed a four

Group	Session	Noise Level	Test Paradigm	Order
GA & GB	LabQuiet	-	P.800	first
Group A	LabNoisyLvl1	37dB(A)	P.800	random
	LabNoisyLvl2	47dB(A)	P.800	
Group B	CSNoisyLvl1	37dB(A)	P.808	random
	CSNoisyLvl2	47dB(A)	P.808	

Table 1: Study setup and levels at which the background noise was reproduced during each test session.

speaker setup as defined in [3], and a “FIREFACE UCX" served as audio interface.

Since we wanted to simulate CS with one of the groups, listeners of Group B were requested to bring their own computer and headphones for the test, which they employed to conduct the last two sessions of the listening test, i.e. “CSNoisyLvl1" and “CSNoisyLvl2". To execute the assessment of the speech files, all participants connected to the Internet to access the same HTML JavaScript based framework [1] , that has been used in multiple CS studies and has been shown to produce good results [4–6].

2.1 Background Noise Signals

The used street noise signal was taken from the background noise database published at [3]. Specifically, we used the 20 seconds long “Outside_Traffic_Crossroads_binaural.wav" audio file. A frequency analysis revealed that most of its energy was concentrated at the low frequencies between 10Hz and 1000Hz. This information we used as parameters for measuring the different levels of noise with the dummy head. This binaural measurement was conducted in dB SPL (sound pressure level) and A-weighted, over the duration of the noise signal.

2.2 Speech Database

The speech stimuli for the listening test were taken from the database number 501 from the ITU-T Rec. P.863 [7] competition, which was kindly provided by SwissQual AG Solothurn, for research purposes. Four Swiss-German speakers were recorded per condition uttering four different sentences in German. A total of 60 stimuli (9s long on avg.) were selected accounting for 15 speech degradation conditions, e.g. send-side ambient background noise of diverse types, white background noise, speech coding at various bitrates, different audio bandwidths (narrowband 300-3400 Hz, wideband 50-7000 Hz, super wideband 50-14000 Hz), and also, combinations of these degradations.

The database also contains subjective quality assessments to the 60 speech stimuli made by 24 different native German listeners, in accordance with the ITU-T Rec. P.800 [1]. 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"). Table 2 provides basic information about the 15 speech degradation conditions, and more details can be found in [7].

[1] https://gitlab.com/zequeira/NoStimuli-SQA.git last accessed March 2020

3. RESULTS

A total of 36 listeners participated in our study and provided 6480 quality scores. Most of them were between 18 and 35 years old. 47.2% female and 52.8% male. All of them came from Germany and 91.7% were German native speakers. Three listeners were not native Germans. They had a very good command of the German language and were therefore allowed to participate in the study. Finally, there were 18 users in both Group A and Group B, and they were randomly assigned to each group.

The collected speech quality ratings from each of the sessions in both groups were analyzed to identify and discard the ratings deemed extreme outliers, i.e. those located at a distance from the median equal or higher than $3.0 \cdot IQR$ (interquartile range) [8]. As a result, a total of 292 ratings were discarded. See Table 3 for a summary. The remaining 6188 ratings were considered for the analysis presented in this work.

3.1 Analysis of Lab vs LabQuiet

The test session “LabQuiet" was the most similar to the conditions at which the Lab-MOS were produced, i.e both were executed with professional audio equipment and without background noise. Thus, to determine the validity of the mean opinion scores (MOS) gathered in the first session (LabQuiet-MOS), we compared it to the Lab-MOS. Then, to analyze the influence of the environment background noise in the speech quality ratings, we contrasted the LabQuiet-MOS against the MOS values gathered in the remaining two sessions. This analysis was made for both Group A and Group B.

A Pearson’s product-moment correlation and the Root Mean Square Error (RMSE) was run to determine the relationship between the ratings collected in laboratory and in “LabQuiet". Strong positive correlation with the Lab-MOS and low RMSE was seen in both groups, $r = 0.97$ ($p < .001$); $RMSE = 0.367$ in Group A and $r = 0.964$ ($p < .001$); $RMSE = 0.423$ in Group B. These results indicates the validity of the collected speech quality scores at the first session in both groups.

Furthermore, we calculated the Pearson’s correlation and RMSE between the laboratory ratings and the rest of the test sessions. As well, a strong positive correlation and low RMSE with the Lab-MOS was observed in all cases, indicating the validity of the speech quality scores collected. These results are outlined in Table 4

3.2 Effect of Background Noise

To assess the influence of the environment background noise on the speech quality scores, listeners executed two more times the listening test under the influence of background noise at two different levels, i.e. 37dBA and 47dBA (see Table 1).

Cond. Number	Description
1	SWB
2	SWB+Noise 12dB
3	SWB+Noise 20dB
6	SWB Level -10dB
7	SWB Level -20dB
32	EFR Live M2L + -20dB + ac. Recording
33	EFR Live M2M + Noise 16dB SNR + Phone NS + DTX UL
43	VoIP WB-Call
44	VoIP WB-Call + -16dB
45	VoIP WB-Call + -8dB
46	VoIP WB-Call + +5dB
47	VoIP WB-Call + Noise 16dB SNR + bad channel + +5dB
48	VoIP WB-Call + Noise 16dB SNR + bad channel + -8dB
49	VoIP WB-Call + Noise 16dB SNR + bad channel + -16dB
50	AAC LC + Noise 14dB SNR + ampl. clipping

Table 2: Labels referring to the speech degradation conditions under test [7].

Session	Removed Ratings
LabQuiet	115
LabNoisyLvl1	31
LabNoisyLvl2	29
CSNoisyLvl1	64
CSNoisyLvl2	53

Table 3: Number of ratings deemed extreme outliers that were discarded in each of the study sessions (292 in total).

Session	r	$RMSE$
LabQuiet (GA)	0.971*	0.367
LabNoisyLvl1	0.967*	0.432
LabNoisyLvl2	0.971*	0.345
LabQuiet (GB)	0.964*	0.423
CSNoisyLvl1	0.980*	0.399
CSNoisyLvl2	0.971*	0.388

*$p < 0.001$

Table 4: Pearson's (r) correlation and root mean squared error ($RMSE$) between the Lab-MOS and the MOS scores collected in all of the test sessions in both Group A (GA) and Group B (GB).

A Friedman test was run to determine if there were differences between the speech quality ratings provided by the listeners at the different test sessions. Pairwise comparisons were performed with a Bonferroni [9] correction for multiple comparisons. We found statistically significant differences between the quality scores provided by the listeners at the different test sessions for 4 (6) of the speech degradation conditions that were under test in Group A (Group B). These results can be seen in Table 5 for Group A and in Table 6 for Group B.

Only 4 of the speech impairments were rated statistically significantly different by the participants in Group A. This is a low number if we consider that there were 15 degradation conditions under test. This outcome suggest that reliable speech quality scores can be collected in the presence of a moderate environment background noise of 47dBA, when conducting the listening test with professional audio equipment.

Cond.	$\chi^2(3)$	p-value	LQ Mdn.	Lvl1 Mdn.	Lvl2 Mdn.	Post Hoc Pairwise Comparison
7	9.898	$= .007$	3.75	3.875	3.375	Lvl1 vs. Lvl2 ($p = .005$)
33	13.765	$= .001$	2.00	2.5	2.375	LQ vs. Lvl1 ($p = .012$) LQ vs. Lvl2 ($p < .001$)
48	8.984	$= .011$	2.00	2.5	2.5	LQ vs. Lvl1 ($p = .008$) LQ vs. Lvl2 ($p = .037$)
50	13.563	$= .001$	1.25	1.00	1.00	LQ vs. Lvl2 ($p = .013$)

Table 5: Speech degradation conditions that were rated statistically significantly different between the different test sessions in Group A. "LQ", "Lvl1" and "Lvl2", correspond to the test session "LabQuiet", "LabNoisyLvl1" and "LabNoisyLvl2", respectively. The "Post Hoc" column presents the results of the pairwise comparisons with Bonferroni correction, showing between which test sessions the speech stimuli were rated differently.

Cond.	$\chi^2(3)$	p-value	LQ Mdn.	Lvl1 Mdn.	Lvl2 Mdn.	Post Hoc Pairwise Comparison
2	6.370	$= .041$	3.125	3.00	3.25	Lvl1 vs. Lvl2 ($p = .030$)
6	9.500	$= .009$	4.75	4.50	4.375	LQ vs. Lvl2 ($p = .010$)
7	6.873	$= .032$	4.25	4.00	3.50	LQ vs. Lvl2 ($p = .024$)
43	7.508	$= .023$	2.875	3.292	3.25	LQ vs. Lvl1 ($p = .020$) LQ vs. Lvl2 ($p < .030$)
48	9.836	$= .007$	2.375	2.375	2.625	LQ vs. Lvl1 ($p = .012$) LQ vs. Lvl2 ($p = .012$)
50	10.138	$= .006$	1.125	1.00	1.00	LQ vs. Lvl1 ($p = .080$) LQ vs. Lvl2 ($p = .080$)

Table 6: Speech degradation conditions that were rated statistically significantly different between the different test sessions in Group B. "LQ", "Lvl1" and "Lvl2", correspond to the test session "LabQuiet", "CSNoisyLvl1" and "CSNoisyLvl2", respectively.

On the other hand, the speech quality scores might be less reliable when collected under the influence of a 47dBA background noise, if participants employ their own computer and headphones for the listening test, as was the case of users in Group B (which is also the case of users in crowdsourcing). In this case 6 speech degradation conditions were rated significantly different among the different test sessions. It should be noted that the first test session (LabQuiet) in Group B was conducted with professional audio equipment, unlike the last two sessions that were conducted with the participants own equipment. This difference in the employed hardware might be one of the reasons for the statistical differences that were seen in this group.

Nevertheless, the number of significant differences is still rather low. And these results are in line with those from [10], where authors collected reliable speech quality scores when listeners conducted the test in the presence of an environmental background noise of 43dBA on average, in a simulated crowdsourcing study. Specifically, listeners

from Group 3 in [10], rated statistically significantly different only 5 of the speech degradation conditions when comparing the assessment test conducted in a silent environment to the one with a 43dBA background noise.

3.3 Effect of Environment (CSNoisy vs LabNoisy)

In this subsection we investigate the differences between the listening test conducted following the Rec. P.800 ("LabNoisy") to the one following the Rec. P.808 ("CSNoisy"), and considering also the effect of the environmental background noise.

To this end, we conducted a two-way mixed ANOVA per degradation condition, to determine if there was an interaction effect between environment ("LabNoisy" vs "CSNoisy") and the level of noise, on the speech quality ratings provided by the listeners. The quality scores were approximately normally distributed, as assessed by visual inspection of a Normal Q-Q Plot. There was homogeneity of variances ($p > .05$) and covariances ($p > .001$), as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. This ANOVA test revealed that there was a statistically significant interaction between the environment and the level of noise on the speech quality scores for only condition number 2, $F(1, 34) = 6.418, p = .016$, partial $\eta^2 = .159$. These results show that almost all speech impairment conditions were similarly rated in "LabNoisy" and in "CSNoisy" test sessions that were conducted under the influence of street background noise.

Moreover, the main effect of noise level provoked a statistically significant difference in the mean quality ratings at the different noise levels for condition 6, $F(1, 34) = 12.254, p = .001$, partial $\eta^2 = .265$ and for condition 7, $F(1, 34) = 17.321, p < .001$, partial $\eta^2 = .337$.

To summarize, the main effect of environment ("LabNoisy" vs "CSNoisy") did not lead to any statistical significant difference in the mean speech quality ratings that were provided by the listeners under the two background noise levels under test. This outcome suggest that in the presence of an environment background noise of level 37dBA to 47dBA, listeners would rate these types of speech degradation conditions similarly, regardless of whether the test is performed in the Lab or in CS.

3.4 Comparison of MOS per conditions

Figure 1 presents a comparison of the MOS values per condition that were given by the listeners in the Lab and in our study sessions in Group A and in Group B. It can be seen that the MOS distribution are quite similar in both graphs, which indicate that listeners in both groups assessed the quality of the speech stimuli similarly, regardless of the employed audio equipment.

The graphs also show that participants tended to overrate the quality of the speech stimuli en the presence of environmental background noise when comparing to the Lab-MOS. This last finding differs partially from the ones in [10] where the presence of environmental background noise did not lead to listeners providing constantly higher or lower quality scores instead, it depended on the speech degradation condition that was under test. However, it is valid to point out that the speech database assessed by the participants in [10] was the number 502 from P.863 [7], whereas listeners in our study assessed the database number 501. In both databases the degradation conditions are similar but clearly the speech stimuli are different. Nevertheless, more investigation would be required to find out about the reasons for these differences.

4. CONCLUSION

This paper investigates the influence of environmental background noise on the speech quality ratings, and the interaction effect between background noise and the test procedure, i.e. traditional P.800 Lab test vs. P.808 CS test. To this end, a study was conducted in the laboratory where listeners judged the quality of speech files in the presence of environmental background noise at two different levels. Participants in Group A conducted part of the test following the Recommendation P.800 [1] ("LabNoisy"), while users in Group B executed the listening test in accordance with the Recommendation P.808 [2] ("CSNoisy").

Our results suggest that reliable speech quality scores could be gathered in the presence of an environmental background noise level of 47dBA. Listeners in our study Group A rated statistically significantly different only 4 degradation conditions out of the 15 that were under test. Whereas in Group B, only 6 conditions were rated significantly different when comparing the MOS scores gathered without noise to the ones collected in the presence of street background noise at 47dBA.

Moreover, a significant interaction effect between the environment and the level of noise was seen for only one speech degradation condition. This result indicates that under the influence of an environmental background noise at levels between 37dBA and 47dBA, listeners would rate the quality of speech files in a similar way despite of the test being conducted according to P.800 (Lab) or P.808 (Crowdsourcing).

5. ACKNOWLEDGMENTS

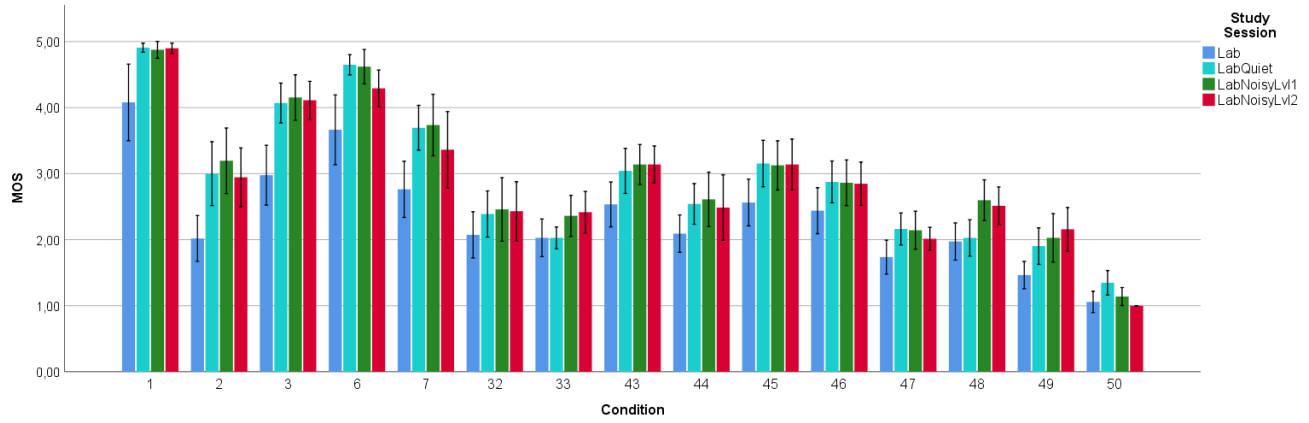
This work has been supported by the German Federal Ministry of Education and Research (BMBF) under Grant No.: FKZ 01IS17052.

6. REFERENCES

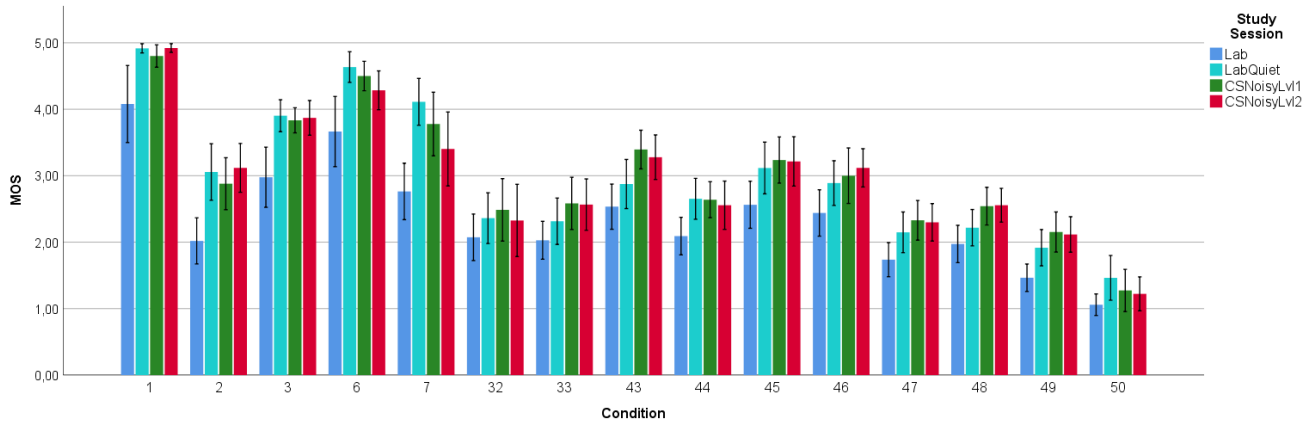
[1] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*. Geneva: International Telecommunication Union, 1996.

[2] ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*. Geneva: International Telecommunication Union, 2018.

[3] E. E. G. . 396-1, *Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance*



(a) Group A



(b) Group B

Figure 1: Comparison between the Lab-MOS and the MOS values per condition collected in the different test sessions with 95% confidence intervals. Information about the degradation conditions can be found in Table 2 and more details in [7].

in the presence of background noise; Part 1: Background noise simulation technique and background noise database. Sophia-Antipolis, France: European Telecommunications Standards Institute, 2011.

[4] R. Zequeira Jiménez, A. Llagostera, B. Naderi, S. Möller, and J. Berger, "Modeling Worker Performance Based on Intra-rater Reliability in Crowdsourcing : A Case Study of Speech Quality Assessment," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2019.

[5] R. Zequeira Jiménez, A. Llagostera, B. Naderi, S. Möller, and J. Berger, "Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing," in *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, (New York, NY, USA), pp. 1138–1143, ACM, 2019.

[6] R. Zequeira Jiménez, L. Fernández Gallardo, and S. Möller, "Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, may 2018.

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

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

[9] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[10] R. Zequeira Jiménez, B. Naderi, and S. Möller, "Effect of Environmental Noise in Speech Quality Assessment Studies using Crowdsourcing," in *2020 Twelve International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.