Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm

Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, Sebastian Möller

Quality and Usability Labs, Telekom Innovation Laboratories, TU-Berlin

{babak.naderi, tim.polzehl, ina.wechsung, friedemann.koester, sebastian.moeller}@telekom.de

Abstract

This paper reports on a crowdsourcing study investigating the influence of trapping questions on the reliability of the collected data. The crowd workers were asked to provide quality ratings for speech samples from a standard database. In addition, they were presented with different types of trapping questions, which were designed based on previous research. The ratings obtained from the crowd workers were compared to ratings collected in a laboratory setting. Best results (i.e. highest correlation with and lowest root-mean-square deviation from the lab ratings) were observed for the type of trapping question, for which a recorded voice was presented in the middle of a random stimuli. The voice explained to the workers that high quality responses are important to us, and asked them to select a specific item to show their concentration. We hypothesize that this kind of trapping question communicates the importance and the value of their work to the crowd workers. Based on Herzberg two-factor theory of job satisfaction, the presence of factors, such as acknowledgment and the feeling of being valued, facilitates satisfaction and motivation, and eventually leads to better performance.

Index Terms: crowdsourcing, speech quality assessment, reliability, motivation

1. Introduction

In human to human communication via telecommunication systems, the quality of transmitted speech as perceived by the user, the so called Quality of Experience (QoE) [1], is assessed by system providers to optimize their services. Traditionally, methods for the QoE assessment of transmitted speech are listening-only-tests (LOTs) [2]. In LOTs, naïve participants are invited to a laboratory and experience multiple stimuli which represent possible degradations of a transmission system. Typically, feedback is gathered on a 5-point Absolute Category Rating (ACR) scale [2]; the average rating for each stimulus over all test participants is called a Mean Opinion Score, MOS. Such lab-based LOTs provide reliable, valid results and are often used as the ground truth for research and industry. However, LOTs conducted in a laboratory setting also exhibit some problematic limitations, as they are: 1) money intensive, concerning the costs for laboratories, participants, test conductors and supervisors, 2) time intensive, concerning invitations and introductions with respect to the number of participants, 3) limited external validity, as laboratory test environments significantly differ from the actual application environment in terms of quality of speech.

Meanwhile, crowdsourcing micro-task platforms offer fast, low cost, and scalable approaches by outsourcing tasks to a large number of users [3], [4]. In addition, crowdsourcing also provides a large diversity of the participants, and a real-life environment for quality assessment of multimedia services and applications [5]. Nevertheless, crowdsourcing user studies cannot be understood as direct implementations of laboratory testing methodologies in an Internet-based environment [5] due to factors they inherit from the nature of crowdsourcing. Hardware and acoustics in the evaluation environment cannot be controlled. Also, results from paid micro-tasks are often noisy due to corrupted responses submitted by participants who are not concentrated enough while working, or who do not work as instructed [3]. For example, instructions may be misunderstood; the crowd workers may be interrupted or may take a break while carrying out the study; or they may split their attention between the study and parallel activities [6]. Moreover in the laboratory experiments, participants mostly continue until the end of the experiment in order to not disappoint the experimenter or to not waste their time while they are already in the lab, but in online crowdsourcing they simply abort the session. Therefore crowdsourcing experiments should be designed differently i.e. oftentimes shorter, more attractive and more unequivocally and transparently structured than the lab experiments. Previous research proposed different approaches such as the use of gold standards, majority voting, and behavioral logging to evaluate results of crowdsourcing experiments in post-processing [7][8]. In the context of QoE, researchers have used additional methods like content questions and consistency tests [9], [10]. The current paper investigates the influence of 'trapping questions', i.e. questions with a known answer, allowing the researcher to identify inattentive or willfully cheating users.

2. Related Work

In [10] authors compare the result of a video quality test in two laboratory studies and different crowdsourcing studies. They found that the correlations between the results from the laboratory and from the unpaid crowd workers (i.e. recruited via social networks) are similar to the correlation between the lab studies. Comparing other conditions (i.e. different monetary rewards and reliability check methods), they conclude that appropriate mechanisms like reliability check and training phases should be included in the crowdsourcing task design. In contrast to the crowdsourcing context, supervision and the face-to-face contact between experimenter and participants in the lab tests may encourage the participants to provide 'good' results [4]. As reported in [3] the presence of an *obvious* reliability check method in a crowdsourcing survey

can significantly increase the consistency of responses. A similar effect is reported by [11]. Here, the authors collected quantitative user ratings and qualitative feedback regarding the quality of Wikipedia articles. They found that if the effort to conceal the act of cheating is as high as the effort to provide reliable answers, respondents are less likely to cheat and eventually the quality of the responses increases.

In both, lab and crowdsourcing context, the motivation of participants is crucial. [3] assumed, that workers may recognize the importance and value of a reliable answer for both, the platform and the job provider, and are thus more motivated to comply with the instructions and refrain from cheating. Similarly, [8] assumed that verifiable questions are signaling to users that their answers will be evaluated. These assumptions are in line with *Herzberg's two factory theory of job satisfaction*. Here, *hygiene factors* and *motivators* are distinguished. Hygiene factors (job context factors such as the working conditions) can in the best case prevent dissatisfaction with the job but they cannot lead to satisfaction [12]. However, their absence will results in dissatisfaction. The absence of motivators (job content factors such as acknowledgment and recognition of the value of the work) on the other hand may not result in dissatisfaction, but their presence will facilitate satisfaction and motivation, which will eventually lead to better performance [12].

With respect to the quality estimation task in crowdsourcing speech assessments, the authors of [13] and [14] suggest that individual web-listeners are well capable of high scores when analyzing their results in terms of absolute and relative differences due to noise level degradations. They defined subgroups of workers according to many characteristics including user characteristics such as age and hearing problems, as well as statistics on so called "anchor-tokens", which they defined as questions of high natural consistency in the seen data, hence offering opportunity for outlier detection. Eventually, they raised the recognition rates from 74% to 87% by selection, at the same time incurring a loss of approx. 75% of the data points. Still, mean correlation (Pearson's r) between lab and crowdsourced data over different noise types and SNR levels resulted in up to 0.8. [15] and 0.96 in [14]. Although these correlations are based on results from intelligibility tests and core of the described studies are not congruent but related to this work, the successful application and the robustness of relative scores can be expected also for the current study on speech quality assessment essential to our treatment on motivation.

3. Research Question

Based on the aforementioned research the current paper inquires whether trapping questions increase the quality of responses collected in a crowdsourcing environment using a LOT quality assessment task. Furthermore we investigate which type of trapping questions work best. Based on motivational theories, we expect that the trapping questions which make participants aware of the importance of their work yield the best results. The second best performance is expected for trapping questions, for which the effort of concealing the cheating would be high, followed by the condition for which the effort of concealing the cheating is low. Poorest outcome is expected for the condition without a trapping task.

We created different types of trapping questions and compared the MOS ratings obtained in the crowdsourcing environment with ratings collected in a lab setting.

4. Method

4.1. Database (SwissQual)

For the experiment, we used stimuli from the database number 501 from the ITU-T Rec. P.863 competition which has been provided by SwissQual AG, Solothurn. This database includes variable types of degradations and degradation combinations. The stimuli were produced according to the ITU-T Rec. P.863 specifications. Overall 200 stimuli are arranged to carry 50 conditions. Each condition describes one degradation or a combination of degradations and each is composed of four stimuli (with the same degradation) recorded by four speakers with four different German sentences. The conditions represent degradations like mixed audio bandwidth (narrowband 300-3400 Hz, wideband 50-7000 Hz, super-wideband 50-14000 Hz), signal-correlated as well as uncorrelated noise, ambient background noise of different types, temporal clipping, speech coding at different bitrates, packet loss of different temporal loss profiles, different frequency distortions, as well as combinations of these degradations. The database contains 24 quality assessments from German natives per stimulus, which were obtained in accordance to ITU-T Rec. P.800. The resulting MOS per stimulus and test condition serves as a reference.

4.2. Crowdsourcing Labeling Procedure

Labeling was processed using the *Crowdee* mobile crowdsourcing platform of the Quality and Usability Lab at Technische Universität Berlin, Germany[1]. Continuing longstanding quality and usability assessments in the labs, this platform provides the opportunity for app-based crowdsourcing assessments including crowdsourcing speech recordings in the field. Workers are mainly students from various places in Germany, many of which have been trained by former lab-studies in the field of HCI. The app to be used by the workers is freely available in the Google Play Store. In principle, everyone can design and run studies on Crowdee. Various filters like, language or gender can be applied to select participants from the crowd. All common media formats are supported for playback and recording. Further custom filters and qualifications can be assigned, which enables automatic job-chains for giving access to specific jobs and training recurrence, as described in the following.

When it comes to study implementation, the research question breaks down into several connected series of micro tasks. Following the interaction flow, the workers were presented three individual jobs: (1) Qualification, (2) Training, (3) Speech quality assessment.

Workers were welcomed by a qualification test of about 2-3 minutes. This test included inquiries on hearing impairment and prior experiences with quality assessment, as well as a technical headset and audio playback test. We asked the workers to seek a quiet place, and controlled the playback volume after asking to adjust it to a comfortable level. We imposed the usage of two-eared headsets such that jobs could only be started when connecting it, and validated its usage by short math exercises with digits panning between left and right in stereo. After approval, workers were assigned to one of the different treatments (i.e. trapping question mechanisms) of the study and automatically given access to a second job, namely

[1] https://www.crowdee.de

training. We presented 12 types of degradations, which are the first 12 anchor stimuli in the SwissQual database, and made the user explicitly aware of the presented degradation types, e.g. added noise, band limitations, packet loss, and different distortion types. Audio-files were pre-loaded on the workers device and could be played multiple times for training. In this test, training expired after one day, and workers were forced to re-train before continuing when the one-day period had expired. Immediately after training, they were given access to the speech quality assessment jobs. The jobs were structured as followed: first being asked to verify the background noise level and to report on their current level of tiredness. After that workers were either presented with five stimuli in case no trapping questions were used or six stimuli (five stimuli plus one trapping stimulus) in case trapping questions were employed. They were asked to rate the quality of each of the non-trapping stimuli. For the trapping stimuli the answer scales may differ (cf. Sec. 4.3). Accordingly, each job contained either 5 or 6 stimuli and workers were free to do up to 40 jobs in a row (i.e. rate all 200 stimuli of the database), or pause in between tasks at their discretion. Workers were forced to listen to the entire stimulus before they could rate it and proceed to the next one. Ratings were collected on a 5-point ACR scale congruent to the scale used in the lab study.

4.3. Trapping Questions

Three different groups representing different configuration for speech-related trapping questions are compared to a control group with no trapping question (*Trapping T0 - No Trapping*). For each studied group, the speech quality assessment jobs were slightly altered by adding one additional stimulus, which was modified. From the original dataset, 40 different stimuli were randomly selected (different speakers, and different degradation conditions) to build a reference trapping stimuli dataset. The dataset was manipulated to create different types of audio trapping questions:

Trapping T1 - Motivation Message: For the first group of trapping stimuli, a message was recorded with a speaker not part of the speech material to be judged. It was appended to the first four seconds of each of the 40 trapping stimuli. The message was as follows: "*This is an interruption. We - the team of Crowdee - like to ensure that people work conscientiously and attentively on our tasks. Please select the answer '' to confirm your attention now*."

In this group, the trapping question was visually identical to the other speech labeling questions, but the trapping stimulus request workers to choose a specific answer option.

Trapping T2 - Low Effort: In the second group of trapping stimuli, one or more animal sounds were inserted in the middle and at the end of each stimulus from the reference trapping stimuli dataset. In this case, the trapping question was also visually different as a different rating scale was used: Workers were asked to indicate whether they recognize the animal sound in the stimulus in a multiple-choice answer format. Here, workers can provide a true answer for the trapping with low effort; full concentration in entire assessment job is not required.

Trapping T3 - High Effort: In the last group of trapping questions, the stimuli, created for the second group (*Trapping 2*), were used but the stimuli were presented together with the ACR rating scale, which was employed for all other stimuli. In addition, a multiple choice question was added at the end of the job (i.e. after being presented with five non-trapping stimuli plus one trapping question). Workers were asked an additional question, they were told to specify all the animal sound(s) that they recognized in any of the previous stimuli. As for *Trapping 2*, the trapping question was visually recognizable, but in case the worker was inattentive while rating the other stimuli, he/she would need to review all previous stimuli again, to find out the correct answer. In this case, the effort to conceal cheating is high.

Trapping T1 emphasize the importance of highly reliable responses to the workers; the objective to employ these kind of trapping questions was to evaluate the hypothesis that participants put more effort in the job if they are aware of the value of their work. With *Trapping T2* and *Trapping T3* we examined the assumption that the likelihood of cheating decreases if the effort to conceal cheating is as high as the effort to accurately complete a task [11].

For all groups, the speech quality assessment questions are shuffled by platform, using the randomization function provided by *Crowdee*.

4.4. Data Collection

The study was conducted using *Crowdee* for 18 days. After submitting an answer to the qualification job, a worker was automatically assigned to one of the four study groups (*T0*, *T1*, *T2*, *T3*) by the platform. As a result, each worker could only perform the speech quality assessment jobs designed for their study group. Overall, 179 workers (87 f., 92 m., M_{age}= 27.9 y., SD_{age}=8.1 y.) participated in the study. Based on the trapping questions, 49 response were rejected (*T1* =2, *T2* =1, *T3* = 46).

5. Results

For all crowdsourcing groups (*T0*, *T1*, *T2*, *T3*), MOS values were calculated for each of the 200 stimuli of the database. Four different Spearman's rank-order correlations, one for each group, were computed to determine the relationship between the MOS ratings obtained from the crowd and the MOS ratings obtained in the lab. Spearman's rank correlation was used as Kolmogorov-Smirnov-test indicated that normality could not be assumed for the data.

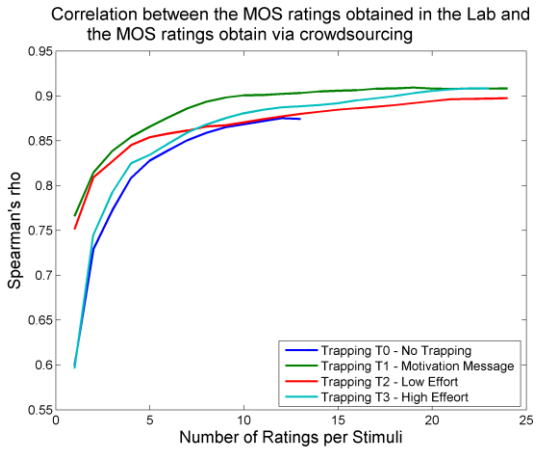
In addition, the Root Mean Square Deviations (RMSD) from the laboratory MOS ratings were calculated for the MOS ratings of the four crowdsourcing groups.

Table 1. *Correlation between the MOS ratings obtained in the lab and the MOS ratings obtained via crowdsourcing (N=200).*

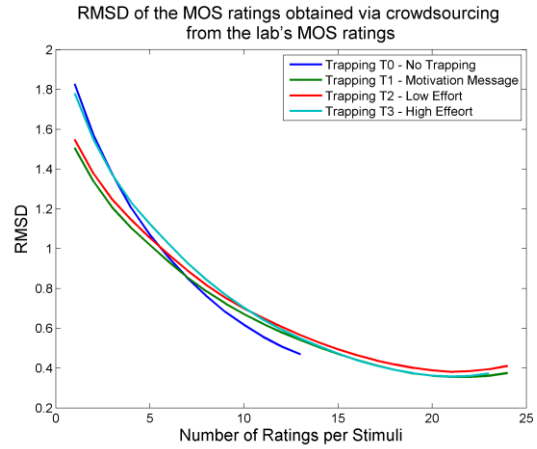
Group	r_s	p-value	RMSD
Trapping T0 - No Trapping	0.886	< 0.001	0.426
Trapping T1 - Motivation Message	0.909	< 0.001	0.375
Trapping T2 - Low Effort	0.897	< 0.001	0.411
Trapping T3 - High Effort	0.909	< 0.001	0.390

Strong positive correlations with the MOS ratings obtained in the lab were observed for all groups, regardless of the kind trapping question employed. However, for *Trapping T1* the highest correlation as well as the lowest RMSD was observed; thus, the results indicate the best performance for this group.

A visual inspection of Figure 1, which displays the changes in correlation and RMSD by the number of participants, confirms these results: To obtain results, which are highly consistent with the lab results (i.e. high correlation and low RMSD) fewer participants are necessary for *Trapping T1*.



(a)



(b)

Figure 1: *Changes in correlation and RMSD depending on number of ratings.*

In the next step, for each of the 50 conditions, we calculated the 95% confidence intervals[1] (CIs) of the mean ratings obtained in the lab and in the crowd. Again the calculations were conducted separately for each group (*T0, T1, T2, T3*).

Based on this data we checked for which and for how many conditions the CIs of the crowdsourcing ratings did not overlap with the CIs of the ratings obtained in the lab. Again the results (cf. Table 2) showed best results for *Trapping 1*: for 35 of the conditions an overlap of the CIs was observed. For both, *Trapping T2* and *Trapping T3*, for 30 conditions the CIs of the means were overlapping with CIs of the lab means. Poorest performance was observed for *Trapping T0.* Next, we examined if the number of overlapping and non-overlapping conditions for *T1*, *T2*, and *T3* differed from the control condition *T0*. A χ^2 - test indicated a statistically significant difference between *T0* and *T1*, χ^2= (1, N = 50)=5.15, *p*=.023.

Accordingly, the results obtained with *Trapping T1 - Motivation Message* are more consistent to lab results, than the results obtained with *Trapping T0 - No Trapping.*

Table 2. *Number of condition with the CIs of the crowd means being lower, higher and overlapping with the CIs of the lab means.*

Group	**N of CIs lower**	**N of CIs higher**	**N of CIs overlapping**
Trapping T0 - No Trapping	17	6	27
Trapping T1 - Motivation Message	13	2	35
Trapping T2 - Low Effort	17	3	30
Trapping T3 - High Effort	16	4	30

In addition, the results show that for 9 of the 50 conditions the MOS in crowdsourcing studies are significantly different from the lab study (≈ 0.5 on the MOS scale). An explorative analysis of the data showed especially narrow-band (NB) speech files tend to be rated with a lower quality in the crowdsourcing study. These results are in line with [17] where NB conditions showed a significant different lower rating. Hence, with respect to the robustness of the quality ratings in crowdsourcing experiments compared to lab studies, the results indicate that some specific condition characteristics provoke a different rating.

6. Discussion and Conclusion

The paper presents a study investigating the influence of different types of trapping questions on the reliability of quality ratings for speech samples obtained in a crowdsourcing environment. Best results were observed for the type of trapping question, for which a recorded voice was presented in the middle of a random stimuli. The voice explained to the workers that high quality responses are important to us, and asked them to select a specific item to show their concentration. A possible explanation for the effect of this kind of trapping questions is that they communicate the importance and the value of their work to the crowd workers. For the other types of trapping questions the effect was weaker; however, also for these questions all obtained data (correlations, RMSD, N of conditions for which the CIs of the means were overlapping with the CIs of the lab means) tended to be more consistent to the lab data compared to the data obtained in the group without any trapping question.

Note that, the *Crowdee* is rather a new platform and the number of jobs is just increasing. Workers might have a different cheating pattern in comparison to other platforms or are not motivated to cheat, yet. Thus, it is likely that the effects of the trapping questions are more pronounced for crowd workers with more experience, something which will be investigated in a future study.

7. Acknowledgements

This project was funded by the Bundesministerium für Bildung und Forschung, Germany (01IS12056) and the Software Campus. The responsibility for the content of this publication lies with the authors.

[1] Note that, CIs offer several advantages compared to p-values [16]: Like p-values, they can be used to estimate the statistical significance of an effect (e.g. non-overlapping 95% CIs indicate a difference on the p<.01 level). Furthermore they can be used to compare different studies and they also provide information regarding the precision of the measure (the wider the CIs the lower the precision of the estimate).

8. References

[1] S. Möller and A. Raake, Eds., *Quality of Experience*. Cham: Springer International Publishing, 2014.

[2] International Telecommunication Union, "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality." Aug-1996.

[3] B. Naderi, I. Wechsung, and S. Möller, "Effect of Being Observed on the Reliability of Responses in Crowdsourcing Micro-task Platforms," in *QoMEX*, 2015.

[4] S. Egelman, E. H. Chi, and S. Dow, "Crowdsourcing in HCI Research," in *Ways of Knowing in HCI*, Springer, 2014, pp. 267–289.

[5] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "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), 1.0, Oct. 2014.

[6] U.-D. Reips, "Standards for Internet-based experimenting.," *Exp. Psychol.*, vol. 49, no. 4, p. 243, 2002.

[7] P. Dai, J. Rzeszotarski, P. Paritosh, and E. H. Chi, "And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions," in *Proc. of 18th ACM CSCW*, 2015.

[8] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini, "Understanding malicious behavior in crowdsourcing platforms: The case of online surveys," in *Proceedings of CHI*, 2015, vol. 15.

[9] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Multimedia (ISM), 2011 IEEE International Symposium on*, 2011, pp. 494–499.

[10] T. Hoßfeld and C. Keimel, "Crowdsourcing in QoE Evaluation," in *Quality of Experience*, Springer, 2014, pp. 315–327.

[11] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing User Studies with Mechanical Turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2008, pp. 453–456.

[12] F. Herzberg, "One more time: how do you motivate employees?," *Harv. Bus. Rev.*, vol. 46, no. 1, pp. 53–62, 1968.

[13] M. K. Wolters, K. B. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using Amazon Mechanical Turk," *SSW*, 2010.

[14] M. Cooke, J. Barker, M. L. G. Lecumberri, and K. Wasilewski, "Crowdsourcing for Word Recognition in Noise.," in *INTERSPEECH*, 2011, pp. 3049–3052.

[15] C. Mayo, V. Aubanel, and M. Cooke, "Effect of prosodic changes on speech intelligibility.," in *INTERSPEECH*, 2012.

[16] G. Cumming and S. Finch, "Inference by eye: confidence intervals and how to read pictures of data," *Am. Psychol.*, vol. 60, no. 2, p. 170, 2005.

[17] T. Polzehl, B. Naderi, F. Köster, and S. Möller, "Robustness in Speech Quality Assessment and Temporal Training Expiry in Mobile Crowdsourcing Environments," *INTERSPEECH* 2015(submitted).