Robustness in Speech Quality Assessment and Temporal Training Expiry in Mobile Crowdsourcing Environments

Tim Polzehl, Babak Naderi, Friedemann Köster, Sebastian Möller

Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin

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

Abstract

Following up on prior work on assessment of quality of speech in laboratory environments, this paper introduces two recently released mobile crowdsourcing paradigms. In comparison to web-based crowdsourcing, mobile crowdsourcing is carried out on smartphones or tablets in the field. Firstly, because involved hardware such as headphones cannot be known in this paradigm, we focus on the effect of mobile crowdsourcing on the assessment of quality of speech using quality degradation types which are described for the model in ITU-T Rec. P.863. As a result, indicators for degradation types that can reliably be assessed in mobile crowdsourcing paradigms are presented for the first time. This reliability is interpreted as robustness towards crowdsourcing assessment environments. Secondly, because working times, pauses and work fragmentation cannot be controlled, we introduce and focus on the analysis of temporarily expiring training certificates as qualifications. Accordingly, we design our study to automatically issue re-training job instances by timeouts, aiming at re-conditioning distracted or oblivious crowd workers. Results indicate a clear improvement in terms of correlation to laboratory test results, when applying the proposed training expiry. Eventually, the indicators presented contribute to build up preliminary guidelines on practical execution of quality assessment using mobile crowdsourcing.

Index Terms: speech quality, mobile crowdsourcing, audio assessment, labeling, annotation, subjective quality

1. Introduction

Various scientific works have focused on the assessment of quality in spoken communication, especially with respect to possible degradations during speech transmission, cf. [1,2]. Oftentimes, these tests are cumbersome, because they require a laboratory environment. Usually 20-30 participants per stimulus to be tested need to be recruited, introduced, and supervised during the tests. Usually, raters then use specific user interfaces (handsets, headphones, etc., of known frequency response characteristics) and carry out their ratings in a quiet neutral environment with certified acoustic (room noise, reverberation time). Raters are given a list of speech files to assess, suggesting or imposing breaks after certain minutes of continuous labor. Every rater receives the same set of stimuli, however in different order. Internal states like inattentiveness or tiredness may arise during imposed overall test time. While raters are usually allowed to interrupt their sessions if they feel tired, rarely any rater leaves early or declares inattentiveness according to our practical experience. Further, it is essentially the artificially imposed "clean" test environment that adversely leads to questionable application in real life, because it can deviate much from the actual application situation that speech is perceived in.

Challenging the above, the ambition to get very accurate quality estimation is oftentimes outweighed by the need to make these studies affordable and practical. Crowdsourcing studies can be carried out at significantly lower costs, but per se do not offer this high level of control and supervision of raters. Crowd workers choose time, duration and pauses at their own discretion, thus causing test fragmentation. They might get distracted and, in case of task-specific assessment training, might forget the trained characteristics and defocus. Distractions stemming from ambient noise and events happening in parallel might have an additional impact on the perceived speech quality. Also, crowd workers bring their own devices, such as smartphones or tablets with unknown characteristics. Eventually, the composition of raters in crowdsourcing is oftentimes much more diverse. Local, temporal and sometimes even cultural differences may cause effect. However, most beneficially crowdsourcing results might have higher external validity and generalize to a larger number of settings [3], and findings are likely to be more applicable to the general population [4].

After introducing the study design in Section 3, we present results from three studies, i.e. overall quality assessment in Section 4.1, the impact of noise (degradation) types in Section 4.2, and the effect of re-train frequency as method to overcome workers' defocus in Section 4.3.

2. Related Work

During the last decade, crowdsourcing has become a fast, low-cost, and scalable opportunity to outsource high volumes of small tasks to a large number of crowd workers, cf. [5,6,7] for a comprehensive overview. In [8] the authors declare, that crowdsourcing, if suitably handled, can be applied for many large labelling tasks, and, crucially, can be comparable to highly paid expert assessments without significant loss in quality. Focusing on quality assessment of an online video service, [9] compares results from two laboratory studies and different crowdsourcing studies. Accordingly, correlations between lab and crowdsourcing studies are similar in magnitude to the correlation between lab studies.

In [10] the authors successfully use both a) naïve crowdsourced; and b) expert workers in laboratory environments to collect human judgments in order to rank the quality of different speech synthesizers. In [11] the authors evaluate the intelligibility of synthesized speech on a purely

crowdsourced basis. Opposed to the present focus on subjective quality assessment (i.e. depending on the listeners' subjective impressions rather than on objectively quantifiable data, such as words recognized), intelligibility is usually assessed in an objective way, e.g. by having listeners reporting on what they have heard out of noise-corrupted speech in written or oral form, which is then wrong or right, and lends itself to calculate word recognition scores. In [12], crowdsourced intelligibility reached laboratory intelligibility by 75% in absolute scores. However, most interestingly the patterns in relative differences remain fairly stable.
Similar finding have been reported in [13], when studying the intelligibility of syllables or words in the presence of noise using crowdsourcing in the so-called *BigListen* case study. Including 600 monosyllabic English words and 12 noise types such as speech-shaped noise, multi-talker bubble noise, and factory noise, the study showed, among many other results, that crowdsourcing can be applied successfully. Note, this study includes listening conditions of different receiver site noise, i.e. quiet room, shared office, Internet café, as well as diverse audio hardware such as headphones, loudspeakers, and laptop speakers. Analyzing the gap in absolute scores, the authors of [11] and [13] suggest that individual web-listeners are well capable of high scores. Selecting subgroups of workers according to a number of specific characteristics, recognition rates could be increased by 12% absolute, while incurring a loss of approx. 75% of the data. Note, previous research proposed different approaches to select, steer and control quality of crowdsourced work by stimulating motivation and/or imposing quality check-ups using various statistics like gold standards, majority voting, or behavioral logging, cf. [14], [15], [16]. For a more detailed discussion on problems of web-based studies cf. [17]. Eventually, mean correlation between lab and crowdsourced data over all noise types and noise SNR results in peaks of 0.8. [12] and 0.96 in [13]. Although these correlations are based on results from intelligibility tests combined with crowd selection methods, the successful application and the robustness of relative scores fosters optimism towards application in speech quality assessment. After all, the foci of the described studies are related but not congruent to the focus of the presented work, as we analyze the applicability of subjective assessment of perceived speech quality in mobile crowdsourcing.

3. Quality Assessment Study Setup

3.1. Laboratory Assessment and Reference Database

Most laboratory tests for assessing the quality of transmitted speech are carried out in a listening-only situation. Here, naïve participants experience a series of stimuli which represent possible degradations of a transmission system. Stimuli are rated either a) individually (e.g. on a 5-point Absolute Category Rating (ACR) scale after ITU-T Rec. P.800 [18], resulting in an averaged Mean Opinion Score (MOS) for each speech file or transmission degradation condition; or b) in pairs (clean vs. degraded speech file) on a 5-point Degradation Category Rating (DCR) scale. In this paper, we concentrate on the ACR ratings, as they are most commonly used for comparing transmission systems.
Stimuli are taken from database number 501 from the ITU-T Rec. P.863 [19] competition which has been provided by SwissQual AG, Solothurn. The database includes various types of degradations and degradation combinations, produced in accordance to the ITU-T Rec. P.863; four speakers with four different German sentences were recorded per condition. Overall, 200 stimuli of German language are arranged to carry 50 degradation conditions, e.g. different audio bandwidths (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 assessments from German natives per stimulus, assessed in accordance to ITU-T Rec. P.800. The respective MOS scores per stimulus and condition serve as reference for further comparisons.

3.2. Crowdsourcing Assessment using Smartphones

Quality assessment was executed using the *Crowdee*[1] scientific mobile crowdsourcing platform, provided by the Quality and Usability Lab, Technische Universität Berlin, Germany. Based on longstanding experiences with quality and usability assessments in laboratory settings, this platform provides the opportunity for mobile-based crowdsourcing assessments. Workers are mainly students from various places in Germany, few have taken part in user studies before. The platform is now also available for non-German participants and the App is freely available in the Google Play Store. All common media formats are supported for playback and recording in the field, including audio assessment.
Further custom filters and qualifications can be assigned, which enable dynamic and automatic crowd-selection and job-orchestration, e.g., giving access to specific jobs as well as training recurrence. Following the course of interaction, the workers are presented three individual jobs:

1. Qualification (one time)
2. Training or re-training (up to 5 times)
3. Speech quality assessment (up to 40 repetitions after each training)

In the *qualification job* workers are asked about hearing impairment or prior experiences with quality assessment. Further, preliminary headset and playback test are commenced. To counteract uncontrolled background noise we ask the workers to seek a quiet place and track the playback volume controller of the device after asking the workers to adjust volume to a comfortable level. We require the workers to use two-eared headsets such that jobs could only be started when connecting headphones, and validate their usage by short math exercises with digits panning left to right in stereo.
After approval, workers are automatically given access to a *training job*, containing 12 stimuli representing 12 anchor types of degradations, which are identical to the first 12 stimuli in SwissQual. During training, we make the workers explicitly aware of the presented degradation types. Speech files are pre-loaded on the workers' device and can be played multiple times during training or assessment. Workers were forced to listen to entire stimuli before being able to assess.
Immediately after training, workers are automatically given access to *speech quality assessment* jobs. After asking to estimate the degree of surrounding background noise level and the temporal degree of tiredness, workers are free to do up to 200 assessment tasks in a row, or pause in between tasks at

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

their own discretion. Stimuli are randomly selected, under the constraint that each worker must not assess any stimulus twice. Assessments are casted on the 5-point ACR scale, in order to be congruent to the laboratory test.

4. Results

In order to provide a common training basis in our first study, training validity expires after one day for all workers, forced re-training on daily basis. Qualification, training and assessment jobs take about 2-3 minutes each, and are rewarded with 30-50 cent. In total, 95 workers (avg=28 years of age, max=54, min=18, std=9, balanced gender) participated in the study, producing 6406 quality assessments, i.e. 24 repetitions for each of the 200 stimuli in the database.

4.1. Comparing Laboratory and Crowdsourcing

When calculating linear correlations according to Pearson between MOS from laboratory assessments and MOS from crowdsourced assessments for individual stimuli, results of 0.92 ($p<0.001$) indicate overall very strong relationship. The corresponding mean-absolute-error (MAE) results in 0.3 points, root-mean-square-error (RMSE) results in 0.38. Figure 1 shows the dependency of ratings in a scatter plot. Standard deviations result in 0.86 and 0.85 for laboratory and crowdsourced assessment, respectively. Consequently, crowdsourcing can be applied to our task with a similar result as in the corresponding laboratory test, i.e. robustly and without sacrificing quality in this respect.

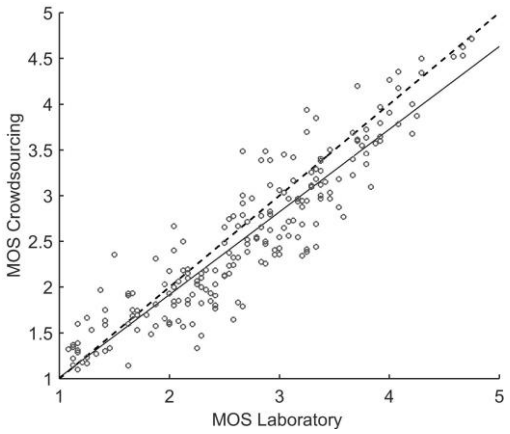

Figure 1: Mean ratings (points) from laboratory and crowdsourcing assessment, perfect (dotted line) and actual (solid line) linear fit.

4.2. Impact of Degradation Types

When analyzing the differences in mean ratings between laboratory and crowdsourced assessments in our second study, individual degradation types can be expected to influence perception in certain ways. The 50 degradation types comprised in the SwissQual database lend themselves to preliminary analyses on significance by applying Tukey's honestly significant difference criterion for post-hoc tests of one-way analyses of variance in between the means of all stimuli that belongs to an individual condition. Results can be summarized as follows:

Audio bandwidth. 19 out of 50 conditions in the SwissQual database are of narrow bandwidth. 17 out of these 19 conditions show a lower MOS for crowdsourced assessment when compared to laboratory assessment; only 2 show a higher MOS. In 5 cases, this difference becomes significant, others show (clear) trends. Consequently, narrow-band speech files seem to provoke a lower quality rating in mobile crowdsourcing than in the laboratory test. In contrast, mean assessments of the 19 wideband and super-wideband conditions do not differ significantly. Thus, wide-band and super-wideband characteristics can be expected to be robust with respect to crowdsourcing assessment.

Bitrate. For 8 conditions containing different bitrates no significant difference was observed in the comparison. Hence, their assessment can be expected to be robust.

Packet Loss. Out of 26 conditions containing packet loss incl. different concealment methods no significant difference is observed. This suggests that also packet loss degradations are robust with respect to mobile crowdsourcing assessment.

Amplitude presentation level. 32 conditions contain amplitude level manipulations in both increased and decreased directions. Results on these conditions are inconclusive, since no systematic behavior can be deduced from the scattered differences in both directions.

Background noise. 12 conditions are recorded with real acoustic background noise at the sending side. Only one condition is assessed significantly lower in crowdsourcing. However, 7 conditions show a trend for lower assessments. Results suggest that crowdsourcing quality assessments comprising acoustic background noise can lead to decreased scores.

4.3. Dynamically Expiring Training Certificates

Finally, the third study focuses on the effect of the temporal distance between training and assessment, and introduces the perspective of dynamically issuing re-training jobs in crowdsourcing paradigms. Essentially, we cannot prescribe the execution of any fixed number of jobs within a certain temporal time span, as crowd workers are free to pause and resume work multiple times. Moreover, a pause could have a positive impact and serve for recovering and re-focusing. On the other hand, it could lead to distraction and defocus. In order to analyze the impact of training and the effect of training recurrence, we use the Crowdee functionality of expiring training certificates as qualification requirement and connect it to our actual assessment job.

Re-Train	N Training	N Worker	Jobs done	Correlation to Labs	MAE
20 min	59	24	1615	0.89	0.36
40 min	50	18	1596	0.89	0.32
60 min	52	20	1595	0.88	0.38
1 day	44	33	1600	0.85	0.41

Table 1: Overview of data, correlations, and mean-absolute-error (MAE) between laboratory and crowdsourced ratings with different training expiry.

We divide workers into 4 non-overlapping groups, as shown in Table 1, following different timeouts of training validity. Upon expiry, raters had to complete a re-training job before

going on with actual assessment jobs. In each group, each stimulus is assessed by at least 8 raters. On average, and regardless of group membership, raters paused 2.2 times for more than 5 minutes, completing 150 tasks and absolving 2.5 training jobs. It is noticeable that the imposed training frequency does not seem to have impact on the work organization pattern with respect to pause and fragmentation.

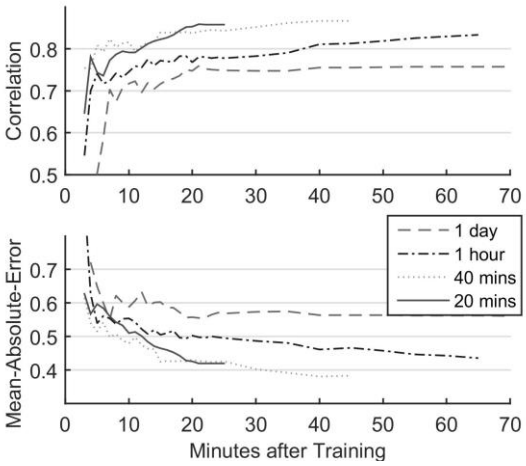

Figure 2: Correlation (top) and mean-absolute-error (bottom) between laboratory and crowdsourced ratings with different training validity expiry.

Figure 2 shows the corresponding correlation to the laboratory assessments when imposing temporally expiring certificates. MOS values are calculated on a temporal grid of 1 minute resolution, not until observing at least three assessments for the respective time span. Correlations are not calculated until at least 5 conditions with calculated MOS are available. Therefore, few data points are present for the first 5-8 minutes roughly, expectedly corrupting the curves during that time. Consequently, the actual correlation calculated need to be interpreted with caution during this time. When analyzing larger time spans, more data leads to more robust statistics.

As a main result, the correlations between labs and crowdsourced assessment saturate earlier and at a higher level when forcing more frequent re-training. This difference becomes significant ($p<0.05$) for the group of 20 and 40 minutes expiry timeout against the group of 1h and 1day expiry. MAEs behave respectively. Table 1 shows the corresponding correlations ($p<0.001$) and error values. The 40 minutes re-train group shows even slightly improvement over the 20 minutes re-train group. As a hypothesis, too frequent training could tear the raters out of concentration and work flow. Hence the optimal training expiry for the current study could be expected to be around 40 minutes. Note, differences in absolute curve length are caused by expiry settings, which principally prevent raters from continuing to work beyond expiry timeout. However, the actual check for training validity happens at the beginning of every task. Curves therefore exceed the expiry timeouts for the actual duration of individual task. Interestingly, continuous work seems to have a kind of training effect in parallel, as results slightly improve along working time for all groups.

5. Conclusion

In this study we show that mobile crowdsourcing can be well suited for speech quality assessment. Using an ITU-T standard speech database incorporating speech degradations we achieve correlations of 0.92, with mean-absolute-error of 0.3 points when compared to assessment under laboratory conditions. In a second study we execute initial experiments in order to analyze the effect of individual speech degradation types on quality assessment in mobile crowdsourcing and laboratory environment. Accordingly, narrow-band speech provokes (partly significantly) lower rating when assessed using crowdsourcing, similar tendencies are observed for background noise. On the other hand, bitrates and packet loss rates are perceived "robustly", i.e. with comparable quality ratings on both sides. Analyzing the effect of work fragmentation with respect to training and defocus from the trained matters by pauses or temporal breaks, we execute a third study on the effect of re-training frequency in order to refresh the cognitive awareness. We introduce the perspective to dynamically impose re-training of crowd workers based on temporal timeouts. We observe that while too frequent training does not improve the results above a certain value of saturation, too infrequent training leads to a decrease of assessment quality. The optimal re-training frequency could empirically be determined to be around 40 minutes for the current tasks and task fragmentation structure.

6. Discussion and Outlook

Oftentimes it is argued, that crowdsourcing may not be suitable for tasks seeking to find out absolute perception performance, but to examine relative differences. Fostering this claim our research shows congruency in relative patterns. Generally, we follow this notion by evaluating mobile crowdsourcing applicability with respect to correlations to laboratory assessments. Absolute assessment tasks might consider crowdsourcing as option for preliminary studies. If in need to raise absolute quality ratings, one should consider to combine the presented methods with crowd selection and motivation procedures. For the current analyzes, advantages of fast and cheap execution are paramount. With respect to amount of data and degradation types in the study, a higher number of samples and a more exclusive stratification of degradation conditions would be desirable, since the SwissQual database does not provide a full-factorial design.
Presumably highly relevant in the paradigm of mobile crowdsourcing using smartphones, our strategy to overcome distractions and effects of task fragmentation is introduced by dynamically expiring training certificates. While we explore the impact of re-training by timeout, a task-dependent estimation of this timeout would be desirable. Also, the number of jobs worked at after training is expected to be highly interesting, as current results suggest that ongoing work also has a training effect. Eventually, the question whether we really need to measure up against laboratory results is crucial in its understanding. Laboratory results can well be expected to deliver "clean" and "single-factored" results. But at the same time one incurs an application scenario mismatch. Especially speech codes and transmission degradation concealment would optimally be designed to deliver best performance in the field, consciously including all potential distortions in everyday life, ultimately welcoming all non-laboratory impacts as relevant factors in evaluation.

7. References

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

[2] S. Egelman, Ed H. Chi, and Steven Dow. "Crowdsourcing in HCI Research," *Ways of Knowing in HCI*. Springer New York, pp. 267-289, 2014.

[3] B. Laugwitz. "A Web experiment on color harmony principles applied to computer user interface design," In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science*, pp. 131–145, Lengerich, 2001.

[4] M.S. Horswill, and M.E. Coster. "User-controlled photographic animations, photograph-based questions, and questionnaires: three Internet-based instruments for measuring drivers' risk-taking behavior," *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society*, Inc, 33(1), pp. 46-58, 2001.

[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)*, 2014.

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

[7] M. Eskenazi, G.A. Levow, H. Meng, G. Parent, D. Suendermann. "Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment," Wiley, 2013.

[8] S. Novotney, and C. Callison-Burch. "Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 207-215, 2010.

[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*, pp. 494–499, 2011.

[10] S. King, and V. Karaiskos. "The Blizzard Challenge 2009," In *Proc. Blizzard Challenge Workshop*, Edinburgh, UK, 2009.

[11] M. K. Wolters, K. Isaac, S. Renals. "Evaluating speech synthesis intelligibility using Amazon Mechanical Turk," *SSW 2010*, pp. 136-141, 2010.

[12] C. Mayo, V. Aubanel, and M. Cooke. "Effect of prosodic changes on speech intelligibility," In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Portland, OR, USA, 2012.

[13] M. Cooke, J. Barker, M.L.G. Lecumberri, and K. Wasilewski. "Crowdsourcing for word recognition in noise," In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,* pp. 3049-3052, 2011.

[14] 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.

[15] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer. "CROWDMOS: An approach for crowdsourcing mean opinion score studies," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2416-2419, 2011.

[16] B. Naderi, I. Wechsung, T. Polzehl, and S. Möller. "Development and Validation of Extrinsic Motivation Scale for Crowdsourcing Micro-task Platforms," *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia.* ACM, pp. 31-36, 2014.

[17] U.D. Reips. "Standards for Internet-based experimenting," *Experimental Psychology*, 49(4), pp. 243-256, 2002.

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

[19] International Telecommunication Union, "ITU-T Recommendation P.863: Perceptual Objective Listening Quality Assessment," 2011.