Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing

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

Anna Llagostera
Rohde & Schwarz SwissQual AG
Zuchwil, Switzerland
anna.llagostera@rohde-schwarz.com

Babak Naderi
Technische Universität Berlin
Berlin, Germany
babak.naderi@tu-berlin.de

Sebastian Möller
Technische Universität Berlin
DFKI Projektbüro Berlin
Berlin, Germany
sebastian.moeller@tu-berlin.de

Jens Berger
Rohde & Schwarz SwissQual AG
Zuchwil, Switzerland
jens.berger@rohde-schwarz.com

ABSTRACT

Crowdsourcing is a great tool for conducting subjective user studies with large amounts of users. Collecting reliable annotations about the quality of speech stimuli is challenging. The task itself is of high subjectivity and users in crowdsourcing work without supervision. This work investigates the intra- and inter-listener agreement withing a subjective speech quality assessment task. To this end, a study has been conducted in the laboratory and in crowdsourcing in which listeners were requested to rate speech stimuli with respect to their overall quality. Ratings were collected on a 5-point scale in accordance with the ITU-T Rec. P.800 and P.808, respectively. The speech samples were taken from the database ITU-T Rec. P.501 Annex D, and were presented four times to the listeners. Finally, the crowdsourcing results were contrasted to the ratings collected in the laboratory. Strong and significant Spearman's correlation was achieved when contrasting the ratings collected in both environments. Our analysis show that while the inter-rater agreement increased the more the listeners conducted the assessment task, the intra-rater reliability remained constant. Our study setup helped to overcome the subjectivity of the task and we found that disagreement can represent a source of information to some extent.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; *Speech / audio search*; • **Human-centered computing** → **User studies**; **Laboratory experiments**; **Computer supported cooperative work**; *Empirical studies in HCI*; • **Computing methodologies** → Speech recognition.

KEYWORDS

inter-rater reliability, speech quality assessment, crowdsourcing, listeners' agreement, subjectivity in crowdsourcing

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317084

ACM Reference Format:
Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller, and Jens Berger. 2019. 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 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317084

1 INTRODUCTION

In the recent years, crowdsourcing has become a convenient approach for solving a multitude of problems that require human input. Crowdsourcing (CS) can be understood as the parallelization of computational work in programming environments, which normally consist of segmenting the work into multiple small and independent tasks. These small tasks, are accomplished by a diverse pool of users in exchange for a monetary compensation. This approach has been adopted in multiple domains and is specially beneficial for conducting subjective user studies.

The quality of the transmitted speech signal is of main importance for telecommunication network providers, as it is one of the main indicators to evaluate their systems and services. The speech signal can be damage by the codecs, linear and non-linear filters, bandwidth limitations, and other elements as reported in [20]. Technological advances within traditional and modern packed-based (Voice-over-IP) telephony networks, introduces new codecs. This demands new empirical subjective user studies to understand how end users perceive these impairments in the speech signal.

Traditionally, subjective speech quality studies has been carried out in Laboratory (Lab) environments under controlled conditions and with professional audio equipment. This way a good control over the experiment setup and a proper control over the participants can be achieved but with some mayor disadvantages, e.g. it is expensive, time consuming, and often the number of participants is rather low. Therefore, sometimes the results might not be representative of a larger population. In contrast, CS permits to reach a fairly demographic distributed pool of users at a fraction of the cost and time.

The study of intra- and inter-rater reliability in speech quality experiments in CS is rather poor. Often, workers evaluate just a portion of the dataset which makes it difficult or even impossible to compute the intra- and inter-rater agreement. This decision

is normally made so the test session can be kept short [7] while avoiding the workers' boredom [21]. This was the case in [18] and in [19], where authors carried a speech quality assessment experiment in a crowdsourcing platform, and workers could assess just the 2.5% of the dataset every time they participated in the study.

Also, work in [25] used CS to investigate the quality of speech stimuli. In the presented experiment, only two workers (out of more than 200) evaluated the entire database and thus, no analysis was given regarding the intra- or inter-rater reliability. Moreover, research in [22] used Amazon Mechanical Turk to investigate the naturalness of synthesized speech in a discrete 5-point scale. Again, workers were able to evaluate just the 3.1% of the available files each time they participated in the study. However, authors repeated the experiment but only to prove the validity of the framework they proposed.

In this work, listeners had the chance to evaluate four times the speech samples in the dataset, so we collected enough data to analyze the agreement among and within subjects. Our hypothesis is that the between and withing listener agreement would increase from the first to the last time the users execute the listening test, as well as the accuracy of the quality scores.

Moreover, authors of [8] proposed the "SOS hypothesis" as an alternative for auditing the reliability of the test results, and to measure consistency in subjective Quality of Experience (QoE) studies. This hypothesis, models a square relationship between the standard deviation of the opinion scores (SOS) and the Mean Opinion Scores (MOS). The authors proposed it also for measuring comparability across multiple QoE studies in CS.

To determine whether our hypothesis is true, we focused on the subjective evaluation of new state-of-the-art codecs (EVS and Opus) in different types of transmission, e.g. VoLTE, circuit switched mobile and VoIP applications like WhatsApp. This evaluation was both carried in the laboratory and in crowdsourcing following international standards, e.g. ITU-T Rec. P.800 [10] and P.808 [12], respectively. Participants in both environments listened four times all of the speech stimuli in the dataset, and gave their opinion about the overall quality in a 5-point scale. The crowdsourcing results are then contrasted to the Lab and we analyze the intra- and inter-listener agreement. Additionally, we investigate the correlation to the Lab results per crowd-worker at the different test session, in order to determine at which one the quality scores are more accurate, which would help to save resources in future studies.

The remainder of this paper is structured as follows. The study setup in the laboratory as well as in the CS environment is outlined in Section 2. Section 3 presents the main findings, firstly the CS results are contrasted with the Lab, and then the agreement between and within subject is analyzed. Section 4 summarizes our findings and concludes.

2 METHOD

In the following, the speech material employed in our investigation is presented as well as the study conducted in the Lab. This latter, aimed at evaluating the user perception of todays' mobile telephony services as VoLTE, circuit switched mobile (GSM and UMTS) and VoIP OTT applications like WhatsApp. Finally, the CS experiment is detailed which intended to replicate the results gathered in the Lab.

2.1 Database

The database consists of a set of real-field and offline samples at different audio bandwidths from below narrowband and up to fullband. The distribution among the speech stimuli is as follows: 21% narrowband (NB), 50% wideband (WB), 23% super-wideband (SWB) and 6% full-wideband (FWB).

The real-field recordings (71% of the speech samples) in the database were collected in Switzerland during September 2018, under good, average and bad coverage network conditions, and using Rohde & Schwarz SwissQual equipment. It includes state-of-the-art measurements such as:

- VoLTE calls with EVS at 24.4 kbit/s SWB
- WhatsApp calls in LTE with Opus at 20 kbit/s WB
- 3G mobile to mobile calls with AMR-WB at 12.65kbit/s and at 23.85 kbit/s
- 3G mobile to mobile calls with AMR-NB at 12.2 kbit/s
- 3G/2G mobile to mobile calls with transcoding from AMR-WB at 12.65 kbit/s to AMR-NB at 12.2 kbit/s

The offline (29% of the speech samples) coded conditions in the dataset aims at replicating those common conditions that can be seen in the field. The three fullband conditions in the test were a fullband reference and two anchors with packet loss. Four additional low quality simulated conditions were obtained by either adding packet loss to some codec conditions, or re-encoding the reference sample multiple times with the same settings. The sample used was the composed female/male German sample from the ITU-T Rec. P.501 Annex D [9]. In total, 53 speech stimuli, 7 seconds long on average, are arranged accounting for 53 degradation conditions.

2.2 Laboratory Study

The study was carried at SwissQual's listening Lab in October 2018. 24 native German listeners (11 female and 13 male) were invited individually to conduct the test. As previously pointed out, this P.800 [10] ACR listening test, targeted the subjective evaluation of speech stimuli encoded with state-of-the-art codecs (e.g. EVS [3], AMR-WB, AMR-NB [1, 2] and Opus [1]) under ideal and live good/average/bad coverage conditions.

Each listener evaluated the quality of the 53 speech stimuli in a five-point absolute category rating (ACR) quality scale with the options (translated from German): "Excellent", "Good", "Fair", "Poor" and "Bad". The presentation order was randomized, and to achieve small enough confidence intervals, each listener evaluated four times the 53 speech samples.

Prior the assessment, participants went through a training in which they judged five stimuli and they become familiar with the test setup. The speech samples were presented diotically to the subjects by means of diffuse field equalized headphones (Grado SR 60). The presentation level was 73 dB(A) SPL at each ear (equivalent to -26dB OVL). More information about the Lab study can be found in [4].

[1] https://opus-codec.org/ last accessed March 2019

All in all, subjective quality assessments to the 53 stimuli were gathered, made by 24 different listeners. The Mean Opinion Score (MOS) [11] was computed for each of the stimulus by averaging the subjective ratings given by all of the subjects to the same speech degradation condition. These MOS scores are then taken as a reference for the analysis presented in this work, from now on referred as “Lab-MOS".

Additionally, a Kendall’s coefficient of concordance (W) [13] was run to determine if there was agreement among the listeners’ ratings. This test revealed a statistical significant agreement across the participants when they assessed all of the speech stimuli, $W = 0.86, p < 0.001$. The reliability of the data is verified when high agreement exists in the ratings that different participants provide to the same speech stimulus. This high Kendall’s coefficient, exposes a low variability across the individual ratings, which demonstrate a high confidence of the collected mean opinion scores.

2.3 Crowdsourcing Study

The goal of the Crowdsourcing (CS) study was to check whether the Lab result could be reproduced in a CS environment. Additionally, we wanted to verify whether the listeners in CS would achieve the same level of agreement considering the subjectivity of the task. With that purpose, we keep constant some important settings from the Lab, e.g. targeting the study to native Germans, gathering at least 96 ratings per sample and conducting the experiment complying with the ITU-T Rec. P.800 [10].

As a CS platform, we used clickworker [2], which is based in Germany as well as most of its users, therefore a good fit for our experimental needs.

The study setup was similar to that one of [25], and with the guidelines of the ITU-T Rec. P.808 [12] in mind, which has been proven to produce good results [17]. Firstly, a screening task was used to collect basic demographic information and to check the workers’ German command, so we could invite to the study only those reporting a native level. For this, a few short German audio-passages were arranged, and the workers were requested to select the right affirmation out of multiple options that were available.

Secondly, the invited workers went through a training phase before they could participate in the speech quality assessment task (SQAT) and evaluate the database. This training permitted to control the headphones’ two-eared usage. A short math exercise was prepared for this with digits panning left to right in stereo [18]. This question allowed to ban the sloppy workers from participating in the SQAT during one hour. Additionally, workers were presented with five stimuli for anchoring, so they could get to know what to expect during the assessment while becoming familiar with the interface. The stimuli employed here were the same as the ones used for anchoring in the Lab study.

Finally, when workers finished the training properly, they were presented automatically with the SQAT. Then, listeners were requested to rate the overall quality of 53 speech stimuli in a five-point scale, see Figure 1. Workers could execute the SQAT up to four times (like in the Lab). Also, one hour timeout was set after which they

[2]https://www.clickworker.com last accessed March 2019

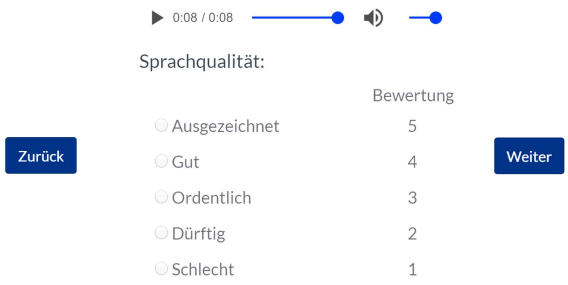


Figure 1: Graphical interface presented to the workers for the SQAT (in line with [12]). The text translate from German: “Speech Quality" and “Rating". The scale (in descending order): “Excellent", “Good", “Fair", “Poor" and “Bad".

were forced to conduct once more the training [19]. To ensure quality, one trapping question (TQ) was inserted randomly within the first five stimuli of every ten speech samples. The TQs’ GUI was the same like in the rest of stimuli, but the audio was modified to highlight the importance of the listeners’ work, and to request them to select an specific option on the rating scale, so they could prove they were conducting the task conscientiously [18]. As result, workers were presented with 58 speech samples in total.

Listeners were prevented from conducting the assessment during one hour, when they missed one of the trapping questions. In this case, the ratings from the set of those ten stimuli were labeled as unreliable. Additionally, at the end of the SQAT workers were requested to express in a slider how tired they felt, 1 was (translated from German): “not exhausted at all" and 11 “extremely exhausted".

To do a “direct" comparison to the Lab results, we wanted to accomplish that at least 24 crowd-workers would execute the SQAT four times, therefore we used a bonus system to motive the participation. Listeners received 1.00 EUR for each time they completed the SQAT, and 0.40 EUR (extra payment) the second, third and fourth time they conducted the SQAT.

3 RESULTS

52 workers in total (100% native Germans, 51.9% female, 96.2% from Germany, balanced age) produced 8321 ratings. Only three workers failed some trapping questions in the SQAT. Two of them missed all of the TQ, while the other worker missed just the last one. Then, a total of 119 ratings were discarded.

To assess whether the ratings gathered in CS correlate to the ones collected in the Lab, we calculated the Spearman’s rank-order correlation. This metric provides a measure of the strength and direction of the association and/or relationship between two continuous or ordinal variables. We then determined the Spearman’s correlation between the Laboratory ratings (Lab-MOS), and the ratings in CS given by the workers that conducted the SQAT four times like in the Lab. 29 workers in total participated four times and yielded 6148 assessments. These subjective ratings were averaged by degradation condition in order to compute the MOS scores, which we refer to as “CS-MOS". Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of

a scatter-plot. Additionally, the Root Mean Square Error (RMSE) between the Lab-MOS and the CS-MOS was calculated. There was a statistically significant, strong positive correlation between the Lab-MOS and the CS-MOS, $\rho = 0.978$ ($p < .001$), as well as a low $RMSE = 0.441$. This result, indicates the validity of CS as a tool for collecting reliable annotations about the quality of speech stimuli, regardless of how close the different degradation conditions may be to each other.

3.1 Inter-rater Reliability

To assess further the validity of the collected ratings, we investigated the level of agreement among the listeners in each of the four stages in which they conducted the SQAT. We wanted to verify whether the agreement fluctuated from the first to the fourth repetition and also if it varied with respect to the Spearman's correlation and the RMSE.

The Kendall's coefficient of concordance (W) [13] can be used as such a measure of inter-rater agreement for continuous and ordinal variables when there are two or more raters [5, 15]. For this analysis we considered the single rating given by the 29 workers to the 53 stimuli at each of the repetitions. Hence, four Kendall's W coefficients were calculated on this ordinal data.

Furthermore, we computed the correlation and RMSE between the Lab-MOS and the ratings provided by the 29 crowd-workers averaged per file, at repetition one, two three and four. Table 1 outline these results. While the RMSE and the correlation remained almost constant, the agreement between workers increased slightly from repetition one to four (e.g. 0.67 to 0.68). This is understandable since a listeners would become more confident with the ratings he or she provides, the more they participate in the study. However, it can be seen that the lowest agreement ($W = 0.65$) was achieved during the third repetition. To investigate further into this, we then conducted a two-way mixed ANOVA [16, 24]. The analysis showed that indeed the main effect of repetition presented a statistically significant difference in the mean ratings at the third repetition, for three of the speech degradation conditions in the dataset, e.g. $C07$, $C15$ and $C37$. Results are presented in Table 2.

We can conclude that the decrease in the between listener agreement at the third repetition, facilitated to identify those three conditions that apparently were the most difficult for the listeners to assess. This outcomes suggest that to gather reliable speech quality annotations for these degradations conditions, the assessment task should be addressed to a large pool of listeners, so the confidence intervals of the quality ratings could be lowered down. In any case, further investigation would be needed to determine the reasons for the differences in the ratings of these speech conditions.

All in all, the high Kendall's coefficient achieved indicate that the SQAT was well designed with a low ambiguity. And we can assume that, all of the workers understood the instructions, and thus, most of the variance across the ratings can be explained by the differences between the speech degradation conditions, and not by individual differences in the listeners' evaluations (e.g. due to different test interpretations). This outcome confirm the reliability of the collected ratings.

Table 1: Kendall's (W) coefficient of agreement, Spearman's correlation (ρ) and RMSE between the Lab-MOS and the CS-MOS for each of the times the workers conducted the SQAT.

Repetition	W	ρ	$RMSE$
1	0.6712^*	0.9774^*	0.4452
2	0.6673^*	0.9714^*	0.4413
3	0.6582^*	0.9750^*	0.4460
4	0.6836^*	0.9780^*	0.4409

$^*p < 0.001$

Table 2: Results of the two-way mixed ANOVA that shows the three conditions for which a statistically significant difference was seen in the mean ratings at the third repetition. Description of these speech degradation conditions are as follow: $C07$ stands for "EVS 13.2 kbps SWB", $C15$ is "2 x AMR-WB 6.6kbps" and $C37$ represent "M2M UMTS call AMR-WB 23.85kbps avg. network conditions 3". More details on the conditions can be found at [4].

Condition	$F(3, 53)$	$p - value$	η^2
$C07$	5.23	$= 0.002$	0.093
$C15$	2.891	$= 0.037$	0.054
$C37$	7.287	< 0.001	0.125

3.2 Analysis of Intra-rater Reliability

Additionally, we investigated at which of the repetitions the workers were more confident with their ratings. With this goal in mind, the intra-rater reliability (IRR) was calculated. The IRR, provides a measure of the consistency in the ratings that a single worker gives to the same sample at different points in time. Then, the IRR was computed for each worker considering the first two, three and four repetitions by calculating the intraclass correlation coefficient (ICC) over the ratings of each SQAT test session [23].

The ICC is a statistical method frequently used for assessing IRR for interval, ratio and ordinal variables, and is especially suitable when the "cases" under research are evaluated two or more times [6], like in our study. Since we were interested in the degree of agreement in the absolute values across the ratings from a single worker, we used a "two-way random" model to compute an $ICC(2, 1)$ coefficient [14]. For this purpose, the *"icc"* function of the R package "IRR" was employed using "agreement" and "single" as parameters [6]. The boxplots in Figure 2 show these results. It can be seen that the repetition did not presented any significant effect on the IRR. This also indicates that all of the workers conducted each of the SQAT with high conscientiousness, which in turn confirms further the reliability of the collected ratings.

Moreover, we explored how the Spearman's correlation (ρ) and the RMSE varied through the whole crowdsourcing study (e.g. from repetition one to four). To this end, we computed per worker and per repetition, the correlation and RMSE between the Lab-MOS and the MOS scores given by a single crowd-worker at each of the test session. Figure 3 presents these results with 95% confidence intervals. It can be seen that both the correlation and the RMSE improved

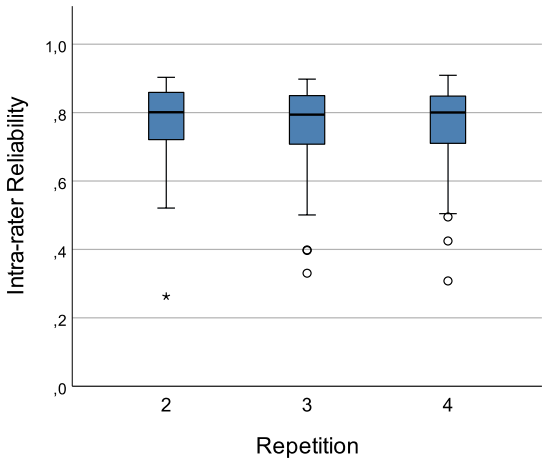


Figure 2: Reliability of the workers at the second, third and fourth time they conducted the SQAT.

with the number of repetitions (e.g. ρ increased from repetition one to four, whereas the RMSE decreased). This outcome evidence that indeed conducting the SQAT multiple times contributed to collect more accurate speech quality scores, and hence better results were achieved after the fourth repetition (when contrasting to the Lab).

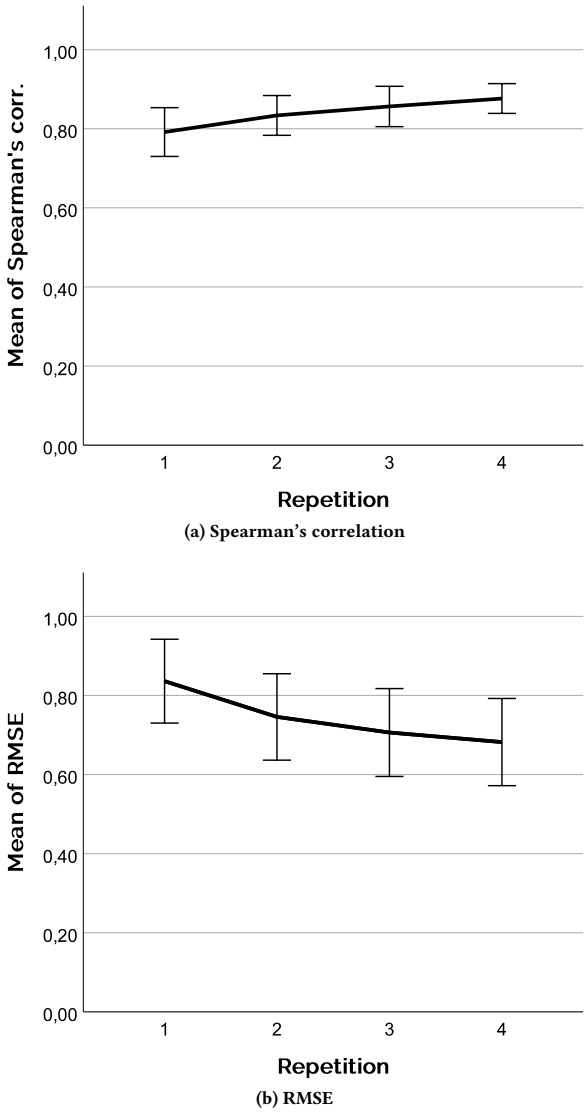
3.3 Discussion

If we remember, our hypothesis was that the inter- and intra-listener agreement would increase gradually from repetition one to four, as well as the accuracy of the speech quality ratings. This later one, in terms of correlation and root-mean-squared-deviation to the Lab results. However, this was not the case with the inter- and intra-rater agreement. While the former fluctuated from the first to the last repetition (see Table 1), the latter remained almost constant as presented in Figure 2. This outcome indicates that there was not a linear relationship between these two metrics, and the fact that listeners as individuals were quite consistent with their answers, it did not contributed to a significant increase in the agreement of the listeners as a group.

Interestingly, an opposite effect was seen with the Spearman's correlation and the RMSE. Both remained almost constant from repetition one to four (see Table 1), when contrasting against the Lab results the ratings of all the workers, whereas both improved with the increase of the repetition number when considering the ratings of individual workers at each of the speech quality assessment test sessions, see Figure 3(a) and 3(b), respectively.

4 CONCLUSION

This work investigate the between and within listeners agreement in a subjective speech quality assessment experiment in crowdsourcing. To this end, a study was conducted in the laboratory in which participants judged the quality of speech files processed with state-of-the-art codecs under different types of transmissions. Ratings were collected in a five-point ACR quality scale in accordance to the ITU-T Rec. P.800 and mean opinion scores were computed. The same study was then replicated in a crowdsourcing platform with



(a) Spearman's correlation

(b) RMSE

Figure 3: Spearman's correlation and root-mean-squared-error (RMSE) with 95% confidence intervals between the ratings provided by each worker and the Lab-MOS.

different participants and a high Spearman's correlation between the Lab-MOS and the CS-MOS was achieved, $r = 0.978$ ($p < .001$).

Our analysis showed that, despite the subjectivity of the task, a proper study setup contributed to collect accurate quality ratings, and the inter-rater agreement increased from the first to the last time the listeners assessed the speech files. However, is worth to point out that a slight decrease in the between listener agreement was seen during the third time the participants conducted the speech quality assessment task. This helped to identify three conditions in the dataset that were quite difficult for the listeners to evaluate. This outcome indicates that disagreement can be also a source of information. Further empirical user studies would be

required to determine the reasons for the differences in the ratings of the aforementioned speech degradation conditions.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 3GPP T S 26.070. [n. d.]. Mandatory speech CODEC speech processing functions; AMR speech Codec; General description.
[2] 3GPP T S 26.171. [n. d.]. Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description.
[3] 3GPP T S 26.441. [n. d.]. Codec for Enhanced Voice Services (EVS); General overview.
[4] Jens Berger and Anna Llagostera. 2018. *A subjective ACR LOT testing super-wideband speech coding in real field measurements and prediction by P.863.* ITU-T Contribution SG12-C.286. International Telecommunication Union, CH-Geneva. 1–11 pages.
[5] Wayne W Daniel. 1980. *Applied nonparametric statistics (2nd ed.).* Boston, MA: Cengage Learning.
[6] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23.
[7] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel. 2014. Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing". https://hal.archives-ouvertes.fr/hal-01078761
[8] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not Enough!. In *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. 131–136. https://doi.org/10.1109/QoMEX.2011.6065690
[9] ITU-T Recommandation P.501. 2017. *Test signals for use in telephonometry.* International Telecommunication Union, Geneva.
[10] ITU-T Recommandation P.800. 1996. *Methods for subjective determination of transmission quality.* International Telecommunication Union, Geneva.
[11] ITU-T Recommandation P.800.1. 2016. *Mean Opinion Score (MOS) Terminology.* International Telecommunication Union, Geneva.
[12] ITU-T Recommandation P.808. 2018. *Subjective evaluation of speech quality with a crowdsourcing approach.* International Telecommunication Union, Geneva.
[13] Maurice George Kendall. 1970. *Rank Correlation Methods* (4th ed.). Charles Griffin.
[14] Richard Landers. 2015. Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS. *The Winnower* (2015). https://doi.org/10.15200/winn.143518.81744
[15] Leonard A Marascuilo and Maryellen McSweeney. 1977. *Nonparametric and distribution-free methods for the social sciences.* Belmont, CA: Wadsworth Publishing Company.
[16] Scott E Maxwell. 1980. Pairwise Multiple Comparisons in Repeated Measures Designs. *Journal of Educational Statistics* 5, 3 (1980), 269–287. https://doi.org/10.3102/10769986005003269
[17] Babak Naderi, Sebastian Möller, and Rafael Zequeira Jiménez. 2018. *Evaluation of the Draft of P.CROWD Recommendation.* ITU-T Contribution SG12-C.204. International Telecommunication Union, CH-Geneva. 1–8 pages. https://www.qu.tu-berlin.de/fileadmin/fg41/publications/naderi{_}2018{_}evaluation-of-the-draft-of-p.crowd-recommendation.pdf
[18] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, and Sebastian Möller. 2015. Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm. In *Interspeech*. ISCA, 2799–2803.
[19] Tim Polzehl, Babak Naderi, Friedemann Köster, and Sebastian Möller. 2015. Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments. In *INTERSPEECH*. 2794–2798.
[20] Alexander Raake. 2007. *Speech Quality of VoIP: Assessment and Prediction.* John Wiley & Sons.
[21] Judith Redi, Ernestasia Siahaan, Pavel Korshunov, Julian Habigt, and Tobias Hoßfeld. 2015. When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal. *Fourth International Workshop on Crowdsourcing for Multimedia* (2015), 33–38. https://doi.org/10.1145/2810188.2810194
[22] Flavio P Ribeiro, Dinei A F Florêncio, Cha Zhang, and Michael L Seltzer. 2011. CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2416–2419. https://doi.org/10.1109/ICASSP.2011.5946971
[23] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin* 86, 2 (1979), 420.
[24] Kevin P Weinfurt. [n. d.]. *Repeated measures analysis: ANOVA, MANOVA, and HLM. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding MORE multivariate statistics.* 317–361 pages.
[25] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. 2018. Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2018.8463298