*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# Subjective Evaluation of Noise Suppression Algorithms in Crowdsourcing

*Babak Naderi*[1], *Ross Cutler*[2]

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]Microsoft Corp., WA, USA

babak.naderi@tu-berlin.de, ross.cutler@microsoft.com

## Abstract

The quality of the speech communication systems, which include noise suppression algorithms, are typically evaluated in laboratory experiments according to the ITU-T Rec. P.835, in which participants rate background noise, speech signal, and overall quality separately. This paper introduces an open-source toolkit for conducting subjective quality evaluation of noise suppressed speech in crowdsourcing. We followed the ITU-T Rec. P.835, and P.808 and highly automate the process to prevent moderator's error. To assess the validity of our evaluation method, we compared the Mean Opinion Scores (MOS), calculated using ratings collected with our implementation and the MOS values from a standard laboratory experiment conducted according to the ITU-T Rec P.835. Results show a high validity in all three scales, namely background noise, speech signal and overall quality (average Pearson Correlation Coefficient (PCC) = 0.961). Results of a round-robin test (N=5) showed that our implementation is also a highly reproducible evaluation method (PCC=0.99). Finally, we used our implementation in the INTERSPEECH 2021 Deep Noise Suppression Challenge [1] as the primary evaluation metric, which demonstrates it is practical to use at scale. The results are analyzed to determine why the overall performance was the best in terms of background noise and speech quality.

**Index Terms**: speech quality, crowdsourcing, P.835, noise suppression, subjective quality assessment

## 1. Introduction

Traditionally, the assessment of the speech quality, transmitted through a telecommunication system, is commonly carried out by human test participants who are either instructed to hold a conversation over a telecommunication system under study (conversation test) or listen to short speech clips (listening-opinion tests) and afterward rate perceived quality on one or several rating scales. Speech calls can be carried out with various devices in different environments, commonly with a non-optimal acoustic surrounding. Therefore, noise suppression algorithms are widely integrated into the communication chain to enhance the quality of the speech communication system. Those systems are typically evaluated in laboratory-based listening tests according to the ITU-T Rec. P.835 [2] in which separate rating scales are used to independently estimate the quality of the *Background noise* (BAK), the *Speech signal* (SIG), and the *Overall quality* (OVRL) alone. Separate scales are used as the higher noise suppression level often adversely affects the speech or the signal component. Consequently, in a regular listening-only test, with a single-rating scale (i.e. according to the ITU-T Rec. P.800), participants can often become confused as to what they should consider in rating the overall "quality." Accordingly, each individual determines their overall quality rating by weighting the signal and background components. Such a process introduces additional error in the overall quality ratings and reduces their reliability [2].

Meanwhile, laboratory-based speech quality experiments are more and more replaced by crowdsourcing-based online tests, which are carried out by paid participants. Crowdsourcing offers a faster, cheaper, and more scalable approach than traditional laboratory tests [3]. Crowdsourcing does have its challenges: the test participants take part in the test in their working environment using their hardware without a test moderator's direct supervision. Previous works showed that background noise in participant's surroundings can mask the degradation under the test and lead to a significantly different rating [4, 5]. Different listening devices can also strongly influence the perceived quality [6]. The ITU-T Rec, P.808 [7] addresses those challenges and provides methods to collect reliable and valid data in the crowdsourcing practice. However, the recommendation only focuses on the Absolute Category Rating (ACR) test method, whereas assessing the noise suppressed speech is more endangered by the environmental noise and uncalibrated listening device. In this work, we followed the methods described in the ITU-T Rec. P.808 and implemented the P.835 test procedure adapted to the crowdsourcing approach. This work's contribution is as follows: we provide an open-source toolkit to conduct subjective assessment of noise suppressed speech in crowdsourcing accessible to the entire research and industry with no need of building a specific laboratory. We show that our approach is highly valid, reliable, reproducible, and scalable in different experiments. Using this tool we are able to provide insights about the state-of-the-art noise suppression algorithms' performance and point out the potential future direction in this domain.

This paper organized as follows: Section 2 describes the toolkit's implementation and different components; Section 3 reports the validity and Section 4 the reproducibility studies we conducted; Section 5 reports the evaluation of different models from INTERSPEECH 2021 Deep Noise Suppression (DNS) Challenge [1] and the relation between the three scales used in the P.835 test. Finally, Section 6 discusses the findings and proposes steps for future work.

## 2. Implementation

We have extended the open-source P.808 Toolkit [8] with methods for evaluating speech communication systems that include a noise suppression algorithm[1]. We followed the ITU-T Rec. P.835 [2] and adapted it for the crowdsourcing approach based on the ITU-T Rec. P.808 [7].

The P.808 Toolkit contains scripts for creating the crowdsourcing test (generate the HTML file, trapping stimuli, input URLs, etc.) and also a script for processing the submitted an-

[1]https://github.com/microsoft/P.808 Accessed June 2021



http://dx.doi.org/10.21437/Interspeech.2021-343

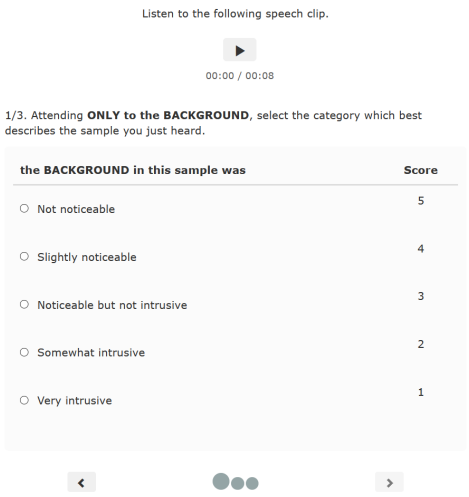


Figure 1: *Screenshot of a trial in P.835 test as presented to the crowd workers.*

swers (data screening and aggregating the reliable ratings). We extended all components to support the P.835 test method. The test includes several sections. In the *qualification* section, relevant demographic questions are asked and the hearing ability of test participants are examined using a digit-triplet test [9]. In the *Setup* section, usage of both ear-pods and the suitability of the participant's environment are evaluated using a modified just-noticeable difference in quality test method [10]. In the *training* section, the test participant is introduced to the rating procedure and familiarized with the rating scale by rating a predefined set of stimuli. The stimuli in the training set should cover the entire range of the scales. The last section is the *ratings* section, in which the participant listens to a set of stimuli and rates them on the given scales. As the crowdsourcing task should be short, therefore, it is recommended to use about ten stimuli in the rating section. The participant can perform one or more tasks. The qualification section only appears once and if the participant passes the test it will not be shown in the next tasks. The setup and training sections appear periodically (every 30 and 60 minutes, respectively) when the worker passed them successfully. We kept the structure of the P.808 Toolkit the same; further details about the P.808 Toolkit can be found in [8] and details on validation of the ITU-T Rec. P.808 in [11].

In the ITU-T Rec. P.835 subjective test, participants are asked to successively attend to and rate the stimulus on the speech signal (from *1-Very distorted* to *5-Not distorted*), the background noise (from *1-Very intrusive* to *5-Not noticeable*) and the overall quality (from *1-Bad* to *5-Excellent*) scales. In our implementation, one clip is used for all three ratings as it was permitted by the recommendation [2]. Participants are forced to listen to the clip again before rating on each scale attending only to the scale's specific aspect[2]. The presentation order of speech signal and background noise scales are randomized in each task to avoid any order effect. The overall quality scale is always the last rating in the sequence. Figure 1 illustrates the P.835 trial as presented to the test participants.

In every crowdsourcing task, it is recommended to include trapping and gold questions [12, 13]. The trapping stimuli is an obvious quality control mechanism [13] which asks the participant to select a specific response to show their attention. For the P.835 extension, the trapping question asks the participant to select a specific score rather than the clip's quality.

Table 1: *Reference conditions for fullband subjective evaluation of noise suppressors according to ETSI TS 103 281 [17].*

| Cond. | Speech Distortion | SNR (A) | Description |
|---|---|---|---|
| i01 | - | - | Best anchor for all |
| i02 | - | 0 dB | Lowest anchor for BAK |
| i03 | - | 12 dB | - |
| i04 | - | 24 dB | - |
| i05 | - | 36 dB | 2nd best anchor for BAK |
| i06 | NS Level 1 | - | Lowest anchor for SIG |
| i07 | NS Level 2 | - | - |
| i08 | NS Level 3 | - | - |
| i09 | NS Level 4 | - | 2nd best anchor for SIG |
| i10 | NS Level 3 | 24 dB | 2nd best anchor for OVRL |
| i11 | NS Level 2 | 12 dB | - |
| i12 | NS Level 1 | 0 dB | Lowest anchor for OVRL |

### 2.1. Reference Conditions

The reference conditions should be used in every subjective test, evaluating noisy speech, to independently vary the signal and background ratings through their entire range of scale values [2]. In the ITU-T Rec. P.835, Signal-to-Noise ratio (SNR) is used for varying the background noise (from 0 to 40 dB) and the Modulated Noise Reference Unit (MNRU) [14] for varying the signal rating (from 8 to 40 dBQ). Overall, 12 reference conditions are recommended. However, a previous study showed that MNRU processing is not appropriate as a reference system for the signal rating scale primary because the degradation in the speech signal by the noise canceller is very different from the degradation resulting from the MNRU processing [15]. Preliminary results from our crowdsourcing test and also expert review showed that software-based MNRU degradation[3] leads to a higher MOS rating than what is reported in the recommendation. Therefore, we applied the twelve reference conditions as proposed in ETSI TS 103 281 [17] (Table 1) in which the *spectral subtraction based distortion* is used for degrading the speech signal. This signal distortion is based on the Wiener filter and leads to similar distortions to the one created by the noise cancellers. We used the configurations as proposed by [17] to create four levels of signal distortions namely *NS Level1* (represents highest signal distortion) to *NS Level 4* (lowest signal distortion). Overall, the reference conditions include: clean speech (i01), conditions with various background noise levels (i02-i05), conditions with various signal distortions (i06-i09), and conditions with both signal and background noise distortions (i10-i12). We used tools provided in [18], clean signals from [19], and noise signals from [17] to create the reference conditions.

[2]Although one time listening reduces the working time on a task, our tests showed that it significantly influences the result.

[3]We used software tools from ITU-T Rec. G.191 [16] to create the processing chain and apply degradations.

## 3. Validation

We conducted a crowdsourcing test using our P.835 extension of the P.808 Toolkit and compared its results with tests conducted in the laboratory according to the ITU-T Rec. P.835. In the test, we used four fullband speech files (2 male, 2 female) from the ITU-T P.501 Annex C [19] and applied the above-mentioned twelve reference conditions on them. On average we have collected 86.2 valid votes per test condition.

Table 2: *Comparison between Crowdsourcing (CS) and Laboratory (Lab) P.835 tests.*

| Scale | PCC CS vs Lab | RMSE CS vs Lab | Average 95% CI CS | Average 95% CI Lab |
|---|---|---|---|---|
| Speech signal | 0.925 | 0.734 | 0.17 | 0.19 |
| Background noise | 0.984 | 0.507 | 0.16 | 0.11 |
| Overall quality | 0.974 | 0.33 | 0.14 | 0.18 |

Figure 2 illustrates the results of the crowdsourcing test. The overall quality ratings tend to be close to the minimum of signal and background noise ratings. Table 2 gives the Pearson Correlation Coefficient (PCC), the Spearman's Rank Correlation Coefficient (SRCC), and their 95% Confidence Intervals (CI). Our results show high correlation to the openly available auditory results conducted in a laboratory from [20].

## 4. Reproducibility study

In [21], the authors carried out four subjective listening tests in three different laboratories to investigate inter- and intra-lab test result repeatability of P.835 methodology. The PCC between tests was high and above 0.97 in all cases.

We used the blind test set from the Deep Noise Suppression (DNS) Challenge [22] and applied 5 DNS models on that test set. We used the outcome and the unprocessed blind dataset for our reproducibility study (i.e. 6 conditions). The blind test set has 700 clips. A P.835 run was done N=5 times, on five separate days, and with mutually exclusive raters for each run. Scatter plot in Figure 3 shows the ratings of six conditions on the three scales for Run1 and Run2 as an example. In Run2 background noise rated slightly higher in most of the conditions but signal and overall quality are rated lower compare to the Run1. Table 3 gives the PCC, SRCC, and the SRCC rank after transformation [23] for N=5 runs, which shows very good reproducibility.

## 5. INTERSPEECH 2021 Deep Noise Suppression Challenge

Our P.835 tool was used in the the INTERSPEECH 2021 Deep Noise Suppression Challenge [1] as the primary evaluation metric. The results of P.835 evaluation for the teams in Track 1 are reported in Table 4. Each team entry was evaluated with $N \sim 2600$ ratings and the resulting 95% confidence interval was 0.04 for MOS values of *BAK*, *SIG*, and *OVRL*.

Table 3: *P.835 reproducibility for N=5 runs. Rank transformation is recommended in case of small number of conditions for Spearman correlation [23].*

| PCC | | | SRCC | | | SRCC trans. rank | | |
|---|---|---|---|---|---|---|---|---|
| OVLR | BAK | SIG | OVLR | BAK | SIG | OVLR | BAK | SIG |
| 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 0.94 | 0.99 | 0.94 | 0.97 |

Hu et al. [24] estimated the below linear relationship between (*BAK*), (*SIG*), and (*OVRL*) using the NOIZEUS dataset:

$$\widehat{OVRL}_{MOS} = -0.0783 + 0.571\, SIG_{MOS} + 0.366\, BAK_{MOS} \quad (1)$$

Using the DNS challenge results and the reference conditions we determined a similar relationship by applying linear regression:

$$\widehat{OVRL}_{MOS} = -0.844 + 0.644\, SIG_{MOS} + 0.452\, BAK_{MOS} \quad (2)$$

This relationship has an adjusted $R^2 = 0.98$ and $\rho = 0.98$, which is similar to the relation [24] estimated with a different dataset. In particular $SIG_{MOS}/BAK_{MOS} = 1.56$ for Eq. (1) and 1.42 for Eq. (2), showing that the signal quality has more weight for overall quality compared to the noise quality.

We applied both Eq. (1) and (2) to the data we collected in Section 3. The predicted values highly correlate with the collected data (PCC=0.97, RMSE=0.51 for Eq. (1), and PCC=0.96, RMSE=0.27 for Eq. (2)), which is shown graphically in Figure 4 for Eq. (2).

The challenge results show that the best overall performing model, *Team 36*, achieves a Difference Mean Opinion Score (DMOS) of 1.01 by very significant noise suppression ($BAK_{DMOS}$=2.05) and with no extra speech degradation ($SIG_{DMOS}$=0.01). There is an additional $BAK_{DMOS}$=0.22 improvement that can be made in noise suppression (comparing to the reference condition i01, i.e., no degradation, with $BAK_{MOS}$=4.88). Using the Eq. (2) for $BAK_{MOS}$=5 and $SIG_{MOS}$=3.89, the predicted $OVLR_{MOS}$ is 3.92, which is an estimate of the best DNS we can do without improving signal quality. To get significantly better performance the speech signal must be improved, such as through dereverberation or capture device distortion removal.

Table 4: *P.835 results from INTERSPEECH 2021 Deep Noise Suppression Challenge Track 1*

| Team | MOS | | | DMOS | | |
|---|---|---|---|---|---|---|
| | BAK | SIG | OVRL | BAK | SIG | OVRL |
| 36 | 4.66 | 3.90 | 3.78 | 2.05 | 0.01 | 1.01 |
| 33 | 4.48 | 3.77 | 3.58 | 1.87 | -0.12 | 0.81 |
| 13 | 4.35 | 3.76 | 3.58 | 1.74 | -0.13 | 0.80 |
| 34 | 4.29 | 3.72 | 3.51 | 1.68 | -0.17 | 0.74 |
| 19 | 4.13 | 3.74 | 3.48 | 1.52 | -0.15 | 0.71 |
| 18 | 4.52 | 3.50 | 3.42 | 1.91 | -0.39 | 0.64 |
| 16 | 3.76 | 3.79 | 3.37 | 1.16 | -0.10 | 0.60 |
| 8 | 4.20 | 3.37 | 3.20 | 1.59 | -0.52 | 0.42 |
| 22 | 4.34 | 3.27 | 3.16 | 1.73 | -0.62 | 0.39 |
| 20 | 3.89 | 3.44 | 3.15 | 1.28 | -0.45 | 0.38 |
| 31 | 3.73 | 3.36 | 3.09 | 1.12 | -0.53 | 0.32 |
| baseline | 3.89 | 3.36 | 3.07 | 1.28 | -0.54 | 0.30 |
| 12 | 4.07 | 3.20 | 3.03 | 1.47 | -0.69 | 0.25 |
| 30 | 3.46 | 3.46 | 2.99 | 0.85 | -0.43 | 0.22 |
| 37 | 4.18 | 3.11 | 2.96 | 1.58 | -0.78 | 0.19 |
| 11 | 3.81 | 3.13 | 2.91 | 1.20 | -0.76 | 0.14 |
| 38 | 2.59 | 3.92 | 2.78 | -0.02 | 0.03 | 0.01 |
| noisy | 2.61 | 3.89 | 2.77 | 0.00 | 0.00 | 0.00 |
| 28 | 3.60 | 2.86 | 2.64 | 1.00 | -1.03 | -0.13 |
| 4 | 2.84 | 3.28 | 2.62 | 0.23 | -0.61 | -0.15 |

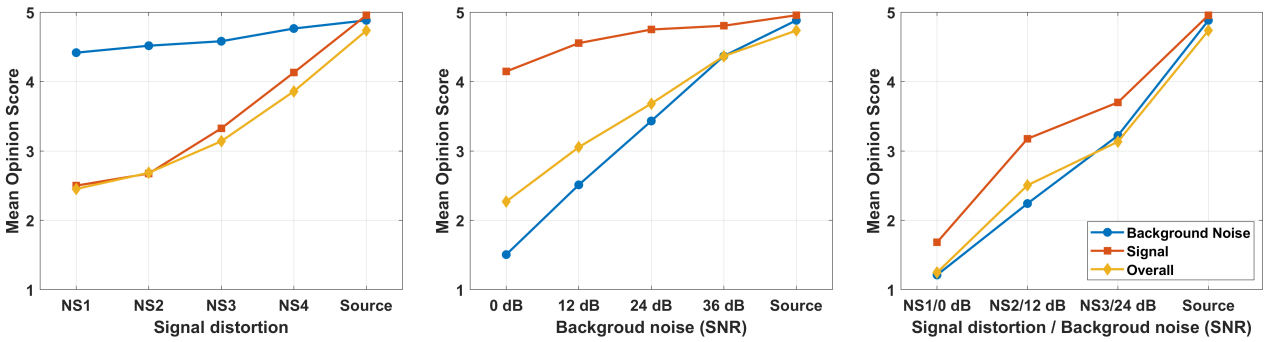


Figure 2: *Auditory results of P.835 CS tests for the reference conditions. (a) Conditions with no background noise, but signal distortion varies, (b) Conditions with background noise varies, but signals not distorted, (c) Conditions with both signal distortion and background noise. Source refers to clean signal (condition i01). NS1-4 refer to different signal distortion levels from Table 1.*

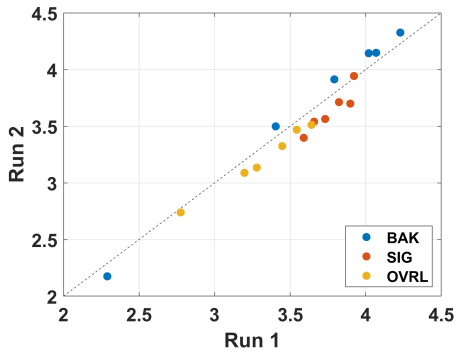


Figure 3: *MOS values of the three scales from two runs in separate days. There are 5 models plus noisy shown.*

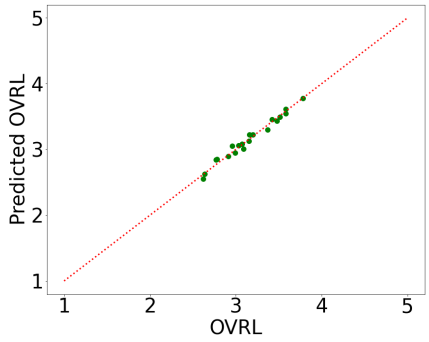


Figure 4: *Linear regression of* $OVRL_{MOS} \sim SIG_{MOS} + BAK_{MOS}$. *Red line is ideal fit.*

## 6. Discussion and Conclusion

We have provided an open-source toolkit for evaluating noise suppression algorithms using subjective tests conducted in the crowdsourcing approach. We followed the ITU-T Rec. P.835 and P.808 and applied the recommended hearing, device usage, and environment suitability tests, as well as the gold standard and trapping questions in the crowdsourcing tasks to ensure the reliability of the estimations. We provide our P.835 implementation as an extension for the P.808 Toolkit. Following the same structure, the toolkit is highly automated to avoid operational errors. The toolkit makes conducting subjective assessment according to the ITU-T recommendations accessible for the entire research and industry community without any extra cost of building the specific laboratory. We conducted a validity study in which we observed high correlations between MOS values of the three scales calculated on ratings collected by our toolkit and the MOS values from the standard ITU-T Rec. P.835 laboratory-based tests (average PCC = 0.961).

We examined the reproducibility of the subjective ratings collected by our implementation. We collected ratings for 4200 clips, including clips processed by five DNS models and the unprocessed ones, in five runs on separate days with mutually exclusive raters. Results show a very good reproducibility (average PCC = 0.98, SPCC = 0.98, SPCC after transformation = 0.97).

We also evaluated the results from the INTERSPEECH 2021 Deep Noise Suppression Challenge [25] using our P.835 implementation. Results show that the best performing model increases the background noise quality by 2.05 MOS without adding extra distortion to the signal quality. Consequently, significant improvement in the performance of the state-of-the-art DNS models can only be achieved by speech signal enhancement. Such a large scale evaluation, in which we collected about 78,000 votes for each rating scale, also demonstrates the scalability of our approach even during the pandemic.

We observed that the P.835 test duration is significantly longer than the P.808 ACR test, as participants need to listen to the speech clips three times. The cost of the P.835 increases by 2× compared to P.808, which is less than 3× due to the common qualification overhead in both P.808 and P.835. For future work, continuous environment monitoring might be considered, which only exposes the participants to a new environment suitability test when a substantial change in the environment is detected. That can significantly decrease the test duration by reducing the extra overhead per session. Meanwhile, the application of continuous rating scales rather than discrete ones should be evaluated considering whether they provide higher sensitivity measurements. We will also evaluate the rating performance as a function of time, which we expect to get worse due to fatigue; perhaps rating fatigue can be detected during the rating process and mitigated using forced breaks in ratings.

# 7. References


[1] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "INTERSPEECH 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[2] ITU-T Recommendation P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, International Telecommunication Union, Geneva, 2003.

[3] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel, "Best practices and recommendations for crowdsourced QoE-lessons learned from the qualinet task force "crowdsourcing"," in *COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET)*.

[4] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller, "Effect of environmental noise in speech quality assessment studies using crowdsourcing," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[5] Babak Naderi, Sebastian Möller, and Gabriel Mittag, "Speech quality assessment in crowdsourcing: Influence of environmental noise," in *44. Deutsche Jahrestagung für Akustik (DAGA)*, pp. 229–302. Deutsche Gesellschaft für Akustik DEGA eV, 2018.

[6] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 2416–2419.

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

[8] Babak Naderi and Ross Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *INTERSPEECH*. 2020, ISCA.

[9] Cas Smits, Theo S Kapteyn, and Tammo Houtgast, "Development and validation of an automatic speech-in-noise screening test by telephone," *International journal of audiology*, vol. 43, no. 1, pp. 15–28, 2004.

[10] Babak Naderi and Sebastian Möller, "Application of Just-Noticeable Difference in Quality as Environment Suitability Test for Crowdsourcing Speech Quality Assessment Task," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[11] Babak Naderi, Rafael Zequeira Jiménez, Matthias Hirth, Sebastian Möller, Florian Metzger, and Tobias Hoßfeld, "Towards speech quality assessment using a crowdsourcing approach: Evaluation of standardized methods," *Quality and User Experience*, vol. 5, no. 1, pp. 1–21, 2020.

[12] Babak Naderi, *Motivation of workers on microtask crowdsourcing platforms*, Springer, 2018.

[13] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, and Sebastian Möller, "Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm," in *Sixteenth annual conference of the international speech communication association*, 2015.

[14] ITU-T Recommendation P.810, *Modulated noise reference unit (MNRU)*, International Telecommunication Union, Geneva, 1996.

[15] AH-11-029, *Better Reference System for the P.835 SIG Rating Scale*, Rapporteur's meeting, International Telecommunication Union, Geneva, Switzerland, 20-21 June 2011.

[16] ITU-T Recommendation G.191, *Software tools for speech and audio coding standardization*, International Telecommunication Union, Geneva, 2019.

[17] ETSI TS 103 281 v1.3.1, *Speech and multimedia Transmission Quality (STQ); Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals* , ETSI, France, 2019.

[18] S4-160397, *Revision of DESUDAPS-1: Common subjective testing framework for training and validation of SWB and FB P.835 test predictors v 1.2* , 3GPP, Memphis, TN, USA, 2016.

[19] ITU-T Recommendation P.501, *Test signals for use in telephony and other speech-based applications*, International Telecommunication Union, Geneva, 2020.

[20] S4-150762, *Reference Impairments for Superwideband and Fullband P.835 Tests – Processing and Auditory Results* , 3GPP, Rennes, France, 2015.

[21] Jan Holub, Hakob Avetisyan, and Scott Isabelle, "Subjective speech quality measurement repeatability: comparison of laboratory test results," *International Journal of Speech Technology*, vol. 20, no. 1, pp. 69–74, 2017.

[22] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[23] Babak Naderi and Sebastian Möller, "Transformation of mean opinion scores to avoid misleading of ranked based statistical techniques," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–4.

[24] Yi Hu and Philipos C Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[25] Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *ICASSP*, 2021.