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

Non-Intrusive Speech Quality Assessment with Transfer Learning and Subject-specific Scaling

Natalia Nessler[1], *Milos Cernak*[2], *Paolo Prandoni*[1], *Pablo Mainar*[2]

[1]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Logitech Europe S.A., Lausanne, Switzerland

natalia.nessler@alumni.epfl.ch, mcernak@logitech.com, paolo.prandoni@epfl.ch, pmainar@logitech.com

Abstract

In communication systems, it is crucial to estimate the perceived quality of audio and speech. The industrial standards for many years have been PESQ, 3QUEST, and POLQA, which are intrusive methods. This restricts the possibilities of using these metrics in real-world conditions, where we might not have access to the clean reference signal. In this work, we develop a new non-intrusive metric based on crowd-sourced data. We build a new speech dataset by combining publicly available speech, noises, and reverberations. Then we follow the ITU P.808 recommendation to label the dataset with mean opinion scores (MOS). Finally, we train a deep neural network to estimate the MOS from the speech data in a non-intrusive way. We propose two novelties in our work. First, we explore transfer learning by pre-training a model using a larger set of POLQA scores and finetuning with the smaller (and thus cheaper) human-labeled set. Secondly, we perform a subject-specific scaling in the MOS scores to adjust for their different subjective scales. Our model yields better accuracy than PESQ, POLQA, and other non-intrusive methods when evaluated on the independent VCTK test set. We also report misleading POLQA scores for reverberant speech.

Index Terms: Speech quality assessment, POLQA, neural networks, transfer learning

1. Introduction

There are various techniques used to evaluate speech quality. The traditional industry standards have been the Perceptual Evaluation of Speech Quality (PESQ) [1], and the Perceptual Objective Listening Quality Assessment (POLQA) [2]. Other methods such as 3QUEST [3] and the Short-Time Objective Intelligibility (STOI) [4] are also widely used. These methods are intrusive: they require a clean reference signal to compare with the signal's noisy version. Moreover, PESQ and POLQA, despite being widely used in various speech assessment tasks, were developed specifically for distortions introduced by speech transmissions such as packet loss, codecs, and compression loss. They were not designed to handle significant reverberations or other common distortions present in modern speech enhancement processing.

There are many applications where having an objective metric for speech quality is crucial: speech denoising and dereverberation [5], computer-assisted language learning [6], speech synthesis [7], and new domains such as speech quality in spatial audio processing. It might be hard if impossible to have access to the clean reference signal in many of them. Therefore the need for non-intrusive objective metrics, where the only input to the model is the distorted signal, is eminent.

Several non-intrusive algorithms were recently proposed to solve the above issues. Microsoft successfully reported using convolutional neural networks for non-intrusive speech quality assessment [8, 9]. In their work, they use crowd-sourcing to label with MOS a speech dataset consisted of 10'000 samples. The Microsoft team achieved robust results compared to intrusive methods such as POLQA, both in root mean square error (RMSE) and in correlation. They also report that PESQ and POLQA scores tend to underestimate the MOS scores. Performing subjective listening tests for ten thousand samples is expensive; could we train accurate models with less data?

Prof. Möller's Quality and Usability Lab also proposed a non-intrusive objective metric in [10]. In their work, they focus on super-wideband speech communication systems. They develop a convolutional network to estimate the per-frame quality and process those with a recurrent network to aggregate them over time. Besides, they made their model publicly available, allowing for comparison with our model. Other works concentrate on predicting speech intelligibility, like in [11]. It is worth noting that speech intelligibility is only one of the aspects of audio quality.

End-to-end models are presented in [12] as a new kind of input for the neural network: raw waveforms are used instead of the more traditional features such as Mel Frequency Cepstral Coefficients (MFCCs). They use PESQ scores as predictors, which according to [8] might not correctly reflect the quality of human perception. Similarly, more recently, the authors of [13] have also developed a non-intrusive metric based on POLQA, PESQ, and STOI labels by using a multi-lingual dataset. As the authors also provide a pre-trained model, we included it in our comparisons.

Until now, training of non-intrusive objective speech quality model has used either labels from existing metrics like POLQA (like in [13, 12]) or has required big datasets of subjective MOS (like in [8]), which are expensive to collect. In this work, we explore a hybrid alternative where we use a relatively small dataset labeled with subjective MOS in addition to a larger dataset labeled with POLQA.

We create a speech dataset by mixing publicly available datasets of speech, noise, and reverberation. We label a subset of this dataset with subjective MOS by doing a crowd-sourcing following the ITU-T P.808 recommendation [14]. The rest of the dataset is labeled with POLQA scores. We evaluate how pre-training a neural net with the larger POLQA datasets impacts the accuracy. We also evaluate a subject-specific scaling of the MOS labels to adjust for their different subjective scales. Finally, we compare with an independent dataset against current industry standards like PESQ or POLQA and other non-intrusive metrics publicly available like [10] and [13].

http://dx.doi.org/10.21437/Interspeech.2021-1685

2. Data generation and subjective rating

We start with the clean samples from LibriSpeech dataset [15], sampled at a rate 16kHz. We choose all the samples approximately 6 seconds long, yielding more than 3000 clean samples, male and female voices distributed equally. For the Room Impulse Responses (RIRs), we use the OpenSLR28 dataset [16]. Various conditions are covered, particularly the size and the shape of the room and the distance to the microphone. The reverberation time (T60) of the room impulse responses covers the range up to 1.4 seconds. We get background noise by randomly drawing a subset of the FreeSound dataset [17], matching the most realistic scenarios like keyboard typing, squeaky chairs, or cafeteria conversations (9 unique noises in total).

To create the dataset, we randomly convolve the clean samples with an impulse of our RIRs subset. We then add the background noise with a random signal-to-noise ratio (SNR) level between -5 and 25 dB. Figure 1 shows the distribution of reverberation time T60 and background noise SNR over the whole dataset. We also keep the clean samples, as well as the samples with reverberation but without noise. The final dataset consists of 30'000 samples (48.9 hours) of speech.

In order to label our dataset, we randomly split it into two parts. The larger part consists of 25'360 samples, and it is evaluated by POLQA on a scale between 1 and 5. On the other hand, the rest of our dataset (4'640 samples) is rated subjectively by human judges on the same scale according to the ITU-T Recommendation Standard P.835 [18] and ITU-T P.808 [14]. For this, Mechanical Turk [19], a crowd-sourcing platform, was used. We used the P.808 toolkit [20] to simplify the experimental design. Each listener needed to provide three labels to each audio sample (without the clean reference, i.e., in a non-intrusive way): (i) the signal's quality, (ii) the background noise's intrusiveness, and (iii) the overall sample quality. We only consider the overall sample quality in all our experiments, and we have not used the other two in this work.

Two hundred eighty-seven listeners rated on average 106 samples each. Various techniques were used to verify the reliability of the judges. The listening test begins with a qualification test, where we verify the eligibility and hearing ability and the listener's environment setup. The listening test continues with four pairs of speech samples; each pair consists of the same sample with the same background noise at two different levels of SNR: 40 [dB] and 50 [dB]. The listener is asked to choose the sample with the best quality in each pair or say that the difference is not detectable. This ensures that the listener can recognize the difference of 10 [dB] in the SNR level and therefore does not use any noise-canceling device that could interfere with our test results.

In the main part of the subjective test, we used gold standard questions to control the quality of the answers. A gold standard is a sample from our dataset with an expected answer. We choose these samples manually to verify that the answer is indeed obvious for all participants. The samples have either very good quality (clean speech without any reverberation or noise) or very bad quality (high T60 and negative SNR). The listener passes this quality check if the given answer is at most one score different from the expected answer: 4 or 5 for the good quality samples and 1 or 2 for the bad ones.

Finally, we used trapping stimuli to motivate the judges to answer the questions attentively. According to [21], the presence of the trapping stimuli in the subjective test improves the quality of the answers. A trapping stimulus is an audio file that begins like any other sample from the dataset; after 2 seconds, the representative sample is interrupted with a motivational message asking to choose a particular answer. The listener passes this quality test if the answer corresponds to the score given in the message.

Besides, listeners with a low overall variance of ratings (less than 0.1) were discarded. A MOS has been computed based on the remaining ratings, using 7.5 scores per sample on average.

We observed a significant variability among different listeners using the range of scores available (from 1 to 5). About 20% of the listeners did not use the entire range of scores. Instead, they rated between 1 and 4, or 2 and 5, or even within a smaller range. In this situation, a single score may have a different meaning depending on the listener and might present a challenge to our neural net that tries to map speech samples into quality scores. To solve it, we scale each listener's ratings to cover the entire range of scores between 1 and 5. Then we compute the resulting MOS for each sample.

In the experimental setup described in the following section, both the original and the scaled MOS are used and evaluated to show the above procedure's impact.

3. Experimental setup

3.1. Training

We randomly split the subjectively rated dataset into training, evaluation, and testing subsets: 70% for training, 20% for validation, 10% for testing.

We standardize the input signal length to a constant value for each file: we use its first 5.5 seconds. The frame and hop lengths are set to 2048 and 512 samples, and Mel-Frequency Cepstral Coefficients (MFCC) computation results in a 13×172 matrix for each sample. We build a neural network consisting of four convolutional layers followed by three fully connected layers. Table 1 shows the details of the proposed network's architecture, inspired by [9]. We use the Rectified Linear activation function (ReLU) in each layer. We use Adam optimizer with learning rate 0.001 and measure the loss using the mean square error (MSE) value.

Table 1: *Network architecture*

Layer	**Parameters**
Convolutional layer	Filters: 16, kernel size: (2, 2)
Batch normalization	—
Max-pooling	Kernel size: (1, 3), strides: (1, 2)
Convolutional layer	Filters: 32, kernel size: (2, 2)
Batch normalization	—
Max-pooling	Kernel size: (3, 3), strides: (2, 2)
Convolutional layer	Filters: 64, kernel size: (2, 2)
Batch normalization	—
Convolutional layer	Filters: 32, kernel size: (2, 2)
Batch normalization	—
Flatten	—
Fully connected layer	Units: 128
Dropout	Rate: 0.5
Fully connected layer	Units: 128
Dropout	Rate: 0.5
Fully connected layer	Units: 1

We explored several candidate models. The first candidate model is trained directly with the dataset with MOS labels (a smaller dataset consisted of 4'640 samples). For the

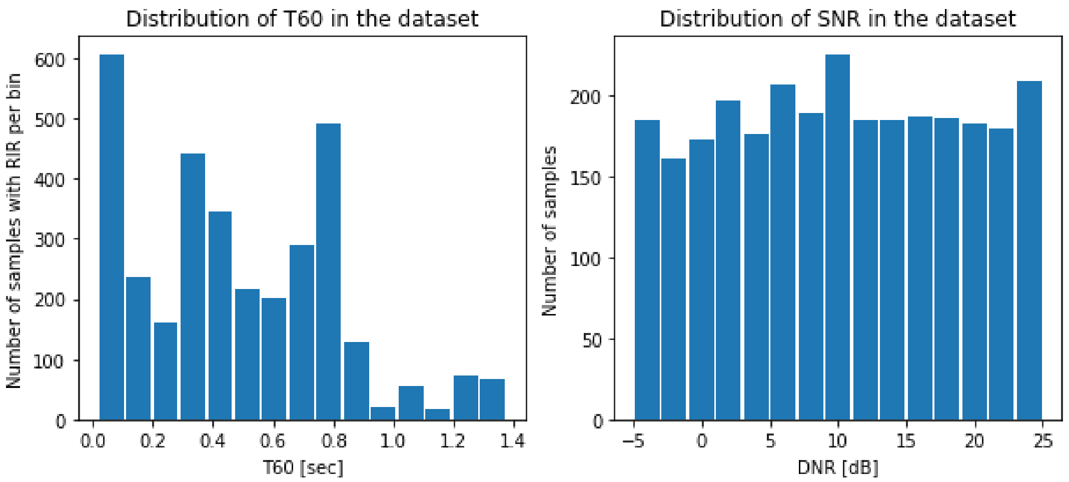

Figure 1: *Distribution of T60 and SNR in the dataset.*

second model, we explored transfer learning. Previous work showed that PESQ scores do not follow the same distribution as MOS [8]. However, we still hypothesized that pre-training with POLQA would initialize the trainable parameters better than random initialization. Thus, we first fully trained a model with the larger dataset (25'360 samples) labeled with POLQA scores. Once it converged, we finetuned all the layers with the smaller dataset labeled with the original MOS. We used the same initial learning rate in both the pre-training and the finetuning because we did not expect the optimal net parameters values to be very close to the values achieved with POLQA, given the findings of [8]. The third and final model was trained in the same manner as the second model, but we used the scaled MOS labels for the finetuning.

3.2. Evaluation

We evaluate the three models using the evaluation dataset labeled with MOS values and compute the Root Mean Square Error (RMSE) and Pearson's correlation (ρ). We also compute the POLQA estimates of the evaluation dataset and compare them against our three models.

Besides, we evaluate our models on an independent open dataset and compare them to other publicly available models. To the best of our knowledge, there is no publicly available dataset with speech quality MOS labels. Therefore, to have a fair comparison with other works, we rely on the work of [22]. The authors ran a subjective test to label the VCTK test dataset and computed the average MOS across all the samples. We process the VCTK test with all the models and find each model's average predicted score. We expect the best one to be the closest to the average showed in [22]. We admit that this is not the ideal model comparison, given that computing the correlation between MOS and model predictions would be more informative. However, without having access to the MOS for each sample, this is the only fair comparison possible.

4. Results

4.1. Comparison of POLQA and subjective MOS

Following on [8] results, we investigate if POLQA scores distribution is different from the MOS obtained in the crowdsourcing experiment. For that purpose, we obtain the POLQA scores of the smaller dataset that is MOS labeled and compare both. The results are shown in Figure 2.

On the bottom of the figure, the samples had no reverberation added to them. For these samples, both MOS and POLQA scores behave very similarly, with values ranging from close to 5 when the SNR is 25 dB to close to 1 when the SNR is -5 dB. However, for samples with reverberation, it can be seen that POLQA scores degrade much faster than MOS scores. Even with a T60 as low as 0.1 seconds, the POLQA scores tend to be very low, almost independently of the SNR.

POLQA was never intended for evaluating reverberant speech. According to ITU-T Recommendation P.863.1 [23], T60 should be below 0.3 seconds above 200Hz, with the distance to the microphone of approximately 10cm. From our findings, we see that POLQA is much more sensible to reverberation than humans are. For example, the MOS of a sample with a reverberation T60 of 0.5 seconds can achieve relatively high values (above 3) even with SNRs as low as 10 dB. On the other hand, POLQA scores for this kind of samples are close to 1.

This mismatch can be problematic for other domains such as speech enhancement. POLQA is the industry standard to measure speech quality, and it can significantly underestimate the MOS for reverberant speech. Therefore it is imperative to have an alternative objective speech quality metric to develop those other domains where reverberation is present.

4.2. Evaluation of the proposed non-intrusive metric

Table 2 shows the RMSE and the correlation of the proposed models and POLQA on our evaluation dataset. As expected, given that the evaluation dataset comes from the same distribution as the training dataset, our models perform better than POLQA. The results confirm that our hypothesis concerning pre-training with POLQA was correct, and we get significantly better results when using transfer learning. These results show that it is possible to develop a non-intrusive metric with a limited amount of MOS-labeled data, which is expensive to collect. By leveraging the easily available POLQA scores, the non-intrusive metric gains accuracy when predicting subjective MOS scores.

Although not directly comparable because of different eval-

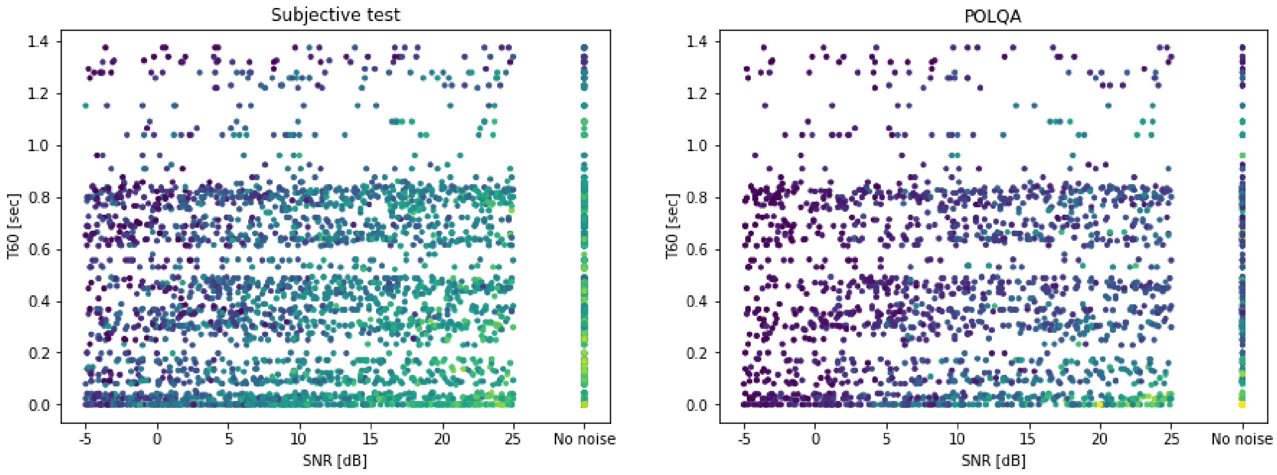

Figure 2: *Comparison of the MOS between the subjective test (left) and POLQA (right). Bright color represents high value, dark color — low value.*

uation dataset is being used, it is interesting to see the results shown by [9]. The authors also split their data into training and evaluation. Their best non-intrusive model reports an RMSE of 0.3742 and a correlation of 0.8792, similar to our results with the pre-trained model. This is despite the fact that their MOS dataset is bigger than ours.

The subject-specific MOS scaling also improves the result, showing that the judges that labeled the MOS dataset may map differently their quality concepts into scores. By doing the subject-specific scaling all the scores are normalized to the maximum range, making the task less challenging for the model.

Table 2: *Results of MOS prediction with respect to MOS.*

Model	RMSE	ρ
POLQA [2]	0.99	0.77
Proposed model 1	0.55	0.85
Proposed model 2 (plus transfer learning)	0.41	0.88
Proposed model 3 (plus MOS scaling)	**0.33**	**0.89**

Additionally, it is interesting to compare our MSE to the MOS variance through our dataset across different listeners. The mean-variance of the MOS is equal to 0.56 in the original dataset and 0.5 in the dataset with scaled MOS. This implies that on average the listeners rate a sample further away from the mean score than our model.

4.3. Comparison to other metrics

The results of the evaluation with the VCTK test set are shown in Table 3. We compare our pre-trained and MOS-scaled model (*proposed model 3*) with industry standards such as PESQ and POLQA. We also compare with other publicly available neural net-based non-intrusive objective metrics like NISQA [10] and WAWENet [13].

The subjective evaluation carried out by [22] with the VCTK test set resulted in an average MOS of 3.40. Our model gets an average MOS of 3.36, which is the closest one compared to other methods. Although lacking the correlation and difficulty to make a significant conclusion, the results indicate that the hybrid approach of pre-training with the POLQA labels and the subject-specific scaling is potentially a good solution for non-intrusive speech quality metrics.

Table 3: *Mean score over the VCTK test set. The proposed model 3 scored the closest to subjective rating.*

Model	Mean score
Subjective evaluation [22]	3.40
PESQ [1]	3.02
POLQA [2]	3.25
NISQA [10]	2.72
WAWENet [13]	3.55
Proposed model 3	**3.36**

5. Conclusions

In this work, we have explored a solution for the expensiveness of collecting subjective MOS data. We have built a hybrid dataset consisting of speech audio partly labeled by human listeners (MOS) and partly labeled by POLQA. By pre-training the model with the cheap-to-collect POLQA scores and then finetuning with the MOS labels, we could achieve the performance of models trained on bigger MOS datasets exclusively. On top of that, performing a subject-specific scaling on the MOS data has been the important pre-processing step to normalize the differences between different human scorers and boosted accuracy and generalization.

Besides, we have explored the insights from [8] about POLQA scores having a different distribution than MOS. We have found that a big part of this mismatch comes from reverberant speech samples. We have highlighted the need for an alternative objective speech quality metric for all the audio domains that work with reverberation, such as speech enhancement. Also, this metric should be non-intrusive to be more useful in real-world conditions where the clean reference signal is rarely available.

6. References

[1] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[2] Y. Gaoxiong and Z. Wei, "The perceptual objective listening quality assessment algorithm in telecommunication: Introduction of itu-t new metrics polqa," in *2012 1st IEEE International Conference on Communications in China (ICCC)*, 2012, pp. 351–355.

[3] ETSI EG 202 393-3, "Speech processing, transmission and quality aspects (stq); speech quality performance in the presence of background noise part 3: Background noise transmission - objective test methods," 2008.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[5] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 53–62, 2018.

[6] A. Bahari, "Computer-assisted language proficiency assessment tools and strategies," *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 36, no. 1, pp. 61–87, 2021. [Online]. Available: https://doi.org/10.1080/02680513.2020.1726738

[7] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the pesq measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[8] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 631–635.

[9] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 85–89.

[10] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7125–7129.

[11] C. Spille, S. Ewert, B. Kollmeier, and B. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech Language*, vol. 48, 10 2017.

[12] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Interspeech*, 2018.

[13] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 331–335.

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

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[16] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[17] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. Interspeech 2019*, pp. 1816–1820, 2019.

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

[19] K. Crowston, "Amazon mechanical turk: A research tool for organizations and information systems scholars," in *Shaping the Future of ICT Research. Methods and Approaches*, A. Bhattacherjee and B. Fitzgerald, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 210–221.

[20] B. Naderi and R. Cutler, "An open source implementation of itu-t recommendation p. 808 with validation," *Proc. Interspeech*, 2020.

[21] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *INTERSPEECH*, 2015.

[22] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," 2020.

[23] ITU-T Recommendation P.863.1, "Application guide for recommendation itu-t p.863," Geneva: International Telecommunication Union, 2019.