ConferencingSpeech 2022 Challenge Evaluation Plan

Gaoxiong Yi[1], Wei Xiao[1], Yiming Xiao[1], Babak Naderi[2], Sebastian Möller[2], Gabriel Mittag[3], Ross Cutler[3], Zhuohuang Zhang[4], Donald S. Williamson[4], Fei Chen[5], Fuzheng Yang[6], Shidong Shang[1]

[1]Tencent Ethereal Audio Lab., China
[2] Technical University of Berlin, Germany
[3] Microsoft Corp., USA
[4] Indiana University Bloomington, USA
[5] Southern University of Science and Technology, China
[6] XiDian University, China

ConferencingSpeech@tencent.com

1. Introduction

With the advances in speech communication systems such as online conferencing applications, we can seamlessly work with people regardless where they are. However, during online meetings, the speech quality can be significantly affected by background noise, reverberation, packet loss, network jitter, and many other factors. Therefore, an effective objective assessment approach is needed to evaluate or monitor the speech quality of the ongoing conversation. If severely degraded speech quality is detected, then the statistical information could be shared with the service provider in time. This helps the service provider to find out the root cause and improve the speech quality to guarantee better user experience in the future. Although the performance of objective speech quality assessment methods has been improved dramatically over the past few decades, there are still a set of open research problems that should be further addressed for the real-time speech communication systems, especially the challenge of non-intrusive speech assessment.

ConferencingSpeech 2022 challenge aims to stimulate research in the areas mentioned above. This challenge will provide comprehensive training datasets, a comprehensive test dataset and a baseline system. The final ranking of this challenge will be determined by accuracy of the predicted MOS from the submitted model or algorithm on the test dataset. We hope this challenge could facilitate idea exchange and discussions in this special session. The tasks in this challenge are in line with speech conferencing applications, which is expected to attract researchers from both academia and industry to participate. In addition, ConferencingSpeech 2022 challenge has the following features:

- This challenge aims to promote the non-intrusive objective quality assessment research for speech communication and targets for effective evaluation on the speech quality of online conferencing applications.
- In recent years, online conferencing applications are getting an increasing amount of attention. Accurate and reliable assessment of speech quality is necessary for the development of those systems. Assessment of speech signals by human listeners, is referred to as subjective speech quality assessment, which is the most preferred approach to assess the speech quality. However, it must be performed under controlled conditions, which is often tedious, time consuming and expensive. Although standardized crowdsourcing tests offer reliable alternatives with strong cost reduction, participants still require compensation, and privacy related concerns limit the usage of this approach. Therefore, objective speech quality assessment methods are needed to effectively evaluate the speech quality in online conferencing applications.
- Different from existing objective speech quality assessment methods, such as Perceptual Evaluation of Speech Quality (PESQ), Perceptual Objective Listening Quality Analysis (POLQA) which need clean reference speech signals as comparison input. This challenge aims to evaluate the speech quality without reference speech signals, which is more practical in online conferencing applications for quality monitoring.
- With the continuous expansion of bandwidth in voice communication systems, the existing standardized non-intrusive objective speech quality assessment method for narrowband speech such as defined in ITU-T P.563 is no longer applicable. Therefore, this challenge aims to effectively evaluate the speech quality for signals with broader bandwidth.
- To truly reflect subjective opinion on speech quality, the training and test datasets used in this challenge no longer adopt PESQ or POLQA score, but the Mean Opinion Score (MOS) as the label, which is obtained through the subjective Absolute Category Ratings (ACR) evaluation in accordance with ITU-T Recommendation P.808.
- As far as we know, this is the first time that the non-intrusive objective speech quality assessment in online conferencing applications is proposed as a challenge. Meanwhile, some additional speech datasets with subjective ratings which were not published before will be shared with the participants. These datasets contain at least 200 hours of speech samples with subjective MOS and covering most of the impairment scenarios users might face in on-line speech communication. It is believed that this will also promote the development of non-intrusive objective speech quality assessment methods.

2. Task Description

This challenge has only one task. In this challenge, comprehensive training and development test datasets with ground truth MOS will be provided to each registered team. More details about the datasets are described in Section 3. It is anticipated that the participating teams using only the impaired speech sig-

Table 1: *Proportion of the pre-processing impaired speech*

Impairment	Percentage
White noise	10%
Nonstationary background noise	60%
High-pass/low-pass filtering	3.75%
Amplitude Clipping	1.25%
AMR/Opus codec	5%
Nonstationary background noise + AMR/Opus codec	5%
White noise + AMR/Opus codec	5%
High-pass/low-pass filtering + AMR/Opus codec	5%
Amplitude Clipping + non-stationary background noise	5%

Table 2: *Impairment scope*

Impairment	Scope1	Scope2	Scope3	Scope4	Scope5
White noise (SNR)	-10 $\sim$ 0dB	0 $\sim$ 10dB	10 $\sim$ 20dB	20 $\sim$ 30dB	30 $\sim$ 40dB
Nonstationary background noise (SNR)	-10 $\sim$ -5dB	-5 $\sim$ 5dB	5 $\sim$ 15dB	15 $\sim$ 25dB	25 $\sim$ 35dB
Low-pass filtering	$<$1000Hz	2400Hz	3600Hz	7200Hz	$>$7200Hz
High-pass filtering	$>$3000Hz	2000Hz	1000Hz	300Hz	$<$300Hz
Clipping	$<$0.02	0.05	0.1	0.4	0.6
AMR codec (rate)	2 $\sim$ 5kb	5 $\sim$ 8kb	8 $\sim$ 15kb	15 $\sim$ 30kb	$>$30kb
Opuscodec (rate)	2 $\sim$ 5kb	5 $\sim$ 8kb	8 $\sim$ 15kb	15 $\sim$ 30kb	$>$30kb

nals to design corresponding algorithms or models, so that the output prediction scores are close to the real MOS. The final ranking of this challenge will be determined by the accuracy of the predicted MOS from the submitted model or algorithm on the evaluation test dataset, in terms of root mean squared error (RMSE).

It is worth noting that there are no restrictions on the source of the training and development test datasets in this challenge. Participants can use any dataset that is beneficial to the designed algorithm or model for development. However, if additional data is used in training, then an ablation study is included that shows the benefit to the test set. Meanwhile, the time-consuming and causality of the proposed algorithm or model are not within the scope of this challenge.

3. Data Description

In this challenge, we provided the participants with four voice datasets along with MOS labels, namely Tencent Corpus, NISQA Corpus, IU Bloomington Corpus, and PSTN Corpus. Among them, except for NISQA Corpus, the other three datasets are both made public for the first time. Each dataset will be described in detail below.

3.1. Tencent Corpus

This dataset includes speech conditions with reverberation and without reverberation. In the without reverberation condition, there are about 10k Chinese speech clips and all speech clips experience the simulated impairments which is very often in online conference. In the with reverberation condition, simulated impairments and live recording speech clips are both considered and totally count about 4k.

3.1.1. Without Reverberation Situation

In the without reverberation condition, the selected source speech clips were artificially added with some damage to simulate the voice impairment scenario that may be encountered in the online meeting scene. The detail of source speech data and simulated impairment is as follows.

3.1.2. Source Data

In order to prevent the possible speaker-dependent behavior of the trained model, the original speech data was selected from three publicly available datasets so as to increase the number of speakers.

- Magic data[1], 50 speakers in total, randomly selected from the original database, reading style, 940 clips, utterance length 5-15s. gender balance 1:1.
- ST Mandarin[2], 855 speakers in total, reading style, 7809 clips, utterance length 2-5s.
- AIshell_100h[3], 184 speakers, reading style, 2056 clips, utterance length 1-5s.

Two utterances were merged into one speech clip to make the final duration of speech clips longer than 5 seconds. The pauses inserted before the first utterance, in the middle of two utterance, and after the last utterance were randomly chosen between 1 to 2 seconds.

3.1.3. Pre-processing

Each speech clip was processed with one type of impairment and only one type, meanwhile there was no same speech utterance with different impairments. The different impairment types and the corresponding percentage of the speech clips applied with each impairment type are listed in Table 1. For each type of impairment, several conditions were considered and listed in Table 2.

3.1.4. The Second Processing Step

Based on the speech clips processed in the first step, we applied another speech processing step including noise suppression and

Table 3: *Proportion of the second step simulated impaired speech*

Impairment	Percentage
Clips processed in the first step and additional noise suppression	10%
Speech clips only processed with the first step impairments	60%
Clean speech	3.75%
Clips processed in the first step and additional noise suppression and packet loss concealment	1.25%
Clean speech and packet loss concealment	1.25%

Table 4: *Network impairments*

Impairment	Percentage
Packet loss	$40\% \sim 70\%$
jitter	$600 \sim 1200$ms
Throttle (bandwidth limitation)	$150 \sim 400$kb

Table 5: *Reverberation parameters*

Room size (m)	Reverberation time (s)
$[5.4, 5.1, 2.7]$	0.4
$[7, 6, 2.7]$	0.5
$[8, 7, 2.8]$	0.6
$[8, 7, 2.8]$	0.7

packet loss concealment to simulate more realistic online communication, and also some clean speech clips were added to form the final speech dataset. Those processing and corresponding percentage in the final dataset are listed in Table 3. For the noise suppression (NS), two algorithms were selected. The first one was the NS algorithm of Tencent meeting (VooV meeting), and the second one was the baseline algorithm of DNS challenge 2021. For the packet loss concealment, the actual speech output was recorded in Tencent meeting client while inputting speech in the far end when facing the network impairments listed in Table 4.

All of those simulation and processing resulted in our final dataset for subjective rating. It contains more than 10k speech clips and the speech length ranges from 5s to 13.5s. Most of the speech clips have length between 5-12s.

3.1.5. Additional Speech Clips with Reverberation

In order to make the subjective database more comprehensive, 4,000 speech clips with reverberation were added to the dataset. 28% of them were generated with simulated reverberation and 72% were recorded in realistic reverberant room. In the simulated reverberation condition, the source data came from the king-asr-166 dataset. Meanwhile, various room sizes and reverberation delays were considered. The specific parameters were shown in Table 5.

Speech clips of daily meetings and conversations in realistic reverberation environment were recorded with microphone placed more than 2 meters away from the speaker. Then the recorded conversation were segmented into speech clips with a random length between 5s and 12s.

3.1.6. Subjective Rating

The subjective scoring procedure was conducted in a crowdsourcing way similar to ITU-T P.808 including qualification - training - rating step, except that the training step was simplified due to the scoring platform we were using. Each clip was rated by more than 24 listeners. After data cleaning more than 20 subjective scores were obtained for each speech clip and averaged to obtain the final MOS score. The distribution of MOS score and 95% CIs have been shown in the following Figure 1 and Figure 2, respectively.

3.2. NISQA Corpus

The NISQA Corpus includes more than 14,000 speech samples with simulated (e.g., codecs, packet-loss, background noise) and live (e.g., mobile phone, Zoom, Skype, WhatsApp) conditions. The corpus is already publicly available so it can only be used as part of the training and development test sets in the competition. Subjective ratings were collected through an extension of P.808 Toolkit in which participants rated the overall quality and the quality dimensions Noisiness, Coloration, Discontinuity, and Loudness. Each clip has on average 5 valid votes. The corpus is organized in 8 datasets:

- NISQA_TRAIN_SIM and NISQA_VAL_SIM: contains simulated distortions with speech samples from four different datasets. Divided into a training set and a validation set.
- NISQA_TRAIN_LIVE and NISQA_VAL_LIVE: contains live phone and Skype recordings with Librivox audiobook samples. Divided into a training set and validation set.
- NISQA_TEST_LIVETALK: contains recordings of real phone and VoIP calls.
- NISQA_TEST_FOR: contains live and simulated conditions with speech samples from the forensic speech dataset.
- NISQA_TEST_NSC: contains live and simulated conditions with speech samples from the NSC dataset.
- NISQA_TEST_P501: contains live and simulated conditions with speech samples from ITU-T Rec. P.501.

For further details about degradations please refer to [4].

In addition, Two evaluation test datasets (each with 200 clips, one in German and one in English) and subjective tests will be conducted using P.808 Toolkit.

3.3. IU Bloomington Corpus

3.3.1. Speech Materials

There are 36,000 speech clips (16-bit single-channel audios sampled at 16 kHz) extracted from COSINE [5] and

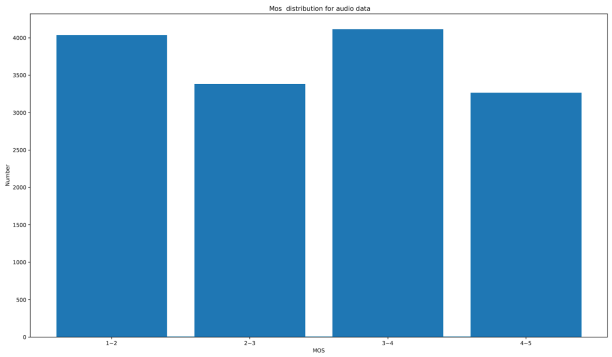

Figure 1: *Distribution of MOS.*

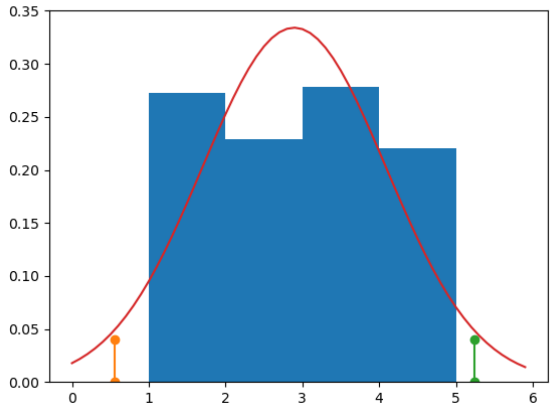

Figure 2: *MOS distribution of 95% CIs.*

VOiCES [6] datasets (16,000 audios each). speech clips are truncated between 3 to 6 seconds long, with a total length of around 45 hours.

For VOiCES dataset, 4 versions of each speech utterance were provided, including reference (i.e., foreground speech), anchor (i.e., low-pass filtered reference), and two reverberant stimuli. The approximated speech-to-reverberation ratios (SRRs) are between -4.9 to 4.3 dB. Three versions of each speech utterance were provided for COSINE dataset, including reference (i.e., close-talking mic), anchor, and noisy (i.e., chest or shoulder mic) stimuli. The approximate signal-to-noise ratios (SNRs) range from -10.1 to 11.4 dB.

3.3.2. Crowdsource Labeling

We crowdsourced our listening tests on Amazon Mechanical Turk by publishing 700 human intelligence tasks (HITs). Each HIT was completed by 5 workers (180k subjective human judgments were collected in total). 3,500 workers (1,455 females and 2,045 males, native English speakers and self-reported to have normal hearing, aged from 18 to 65 years old) participated in the online listening study. Each HIT contains 15 trials of evaluations that follow ITU-R BS.1534 [7]. Each trial has multiple stimuli from varying conditions including a hidden clean reference, an anchor (low-pass clean reference) and multiple real-world noisy or reverberant signals. Participants provided quality ratings (between 0 to 100) for all stimuli. A rescaling step was then performed to normalize the rating ranges (Min-max normalization between 0 to 10). For more details, please refer to [8].

3.4. PSTN Corpus

3.4.1. Speech Materials

The clean reference files used for the phone calls are derived from the public audiobook dataset Librivox. The Librivox corpus contains recordings of 11,350 volunteers reading public domain audiobooks. Because many of the recordings are of poorer quality, the files have been filtered according to their quality as described in [9], leaving in a total 441 hours from 2150 speakers of good quality speech. These audiobook chapters were then segmented into 10 seconds clips and filtered for having a speech activity of at least 50%. Since, in practice, there are often environmental sounds present during phone calls, we used the DNS Challenge 2021 [9] to add background noise. The noise clips are taken from Audioset [10], Freesound, and the DEMAND [11] corpus and added to the clean files with an SNR between $0 - 40$ dB.

Overall, we conducted more than half a million automated phone calls between a PSTN and a VoIP end-point. Because most of these calls were of good quality, we sampled a subset by putting less weight on files with a high POLQA MOS, while maintaining clip and provider diversity. As a result, for the training set 58,709 degraded speech clips with a duration of 10 seconds are available, with 40,739 files based on noisy reference files, and 17,970 files based on clean reference files. The test set consists of 3,000 files, where 2,200 files are based on noisy reference files and 800 files are based on clean reference files.

3.4.2. Crowdsource Labeling

The perceived speech quality of the training and test sets were annotated in a listening experiment on AMT, according to P.808. Each training set file was rated by 5 participants, while the test set files were rated by 30 participants to ensure a low confidence interval of the MOS values for the model evaluation. The participants of all experiments were presented with the same six training files that cover the full quality range. Before calculating the MOS values, the ratings were screened against outliers and unexpected behavior from the crowdworkers. As a consequence, the resulting number of ratings of an individual file may be less than 5 for the training set or 30 for the test set, depending on the screening. For more details, please refer to [12].

3.5. Dataset Division

The training, development, and evaluation test sets in this challenge are all originated from the above-mentioned datasets. It is worth noticing that differs from Tencent, NISQA and PSTN corpora that used ITU-T P.808 for subjective testing, the IU Bloomington corpus adopted ITU-R BS.1534 for subjective test, which resulted in a rating range of $0\sim100$ instead of $1\sim5$. Thus, the IU Bloomington corpus will only be provided to participants as additional materials, speech clips from IU Bloomington corpus will not appear in the evaluation test set of the challenge. Participants can decide whether to use it according to their needs.

Due to the imbalanced size of the datasets, 80% of Tencent Corpus and 95% of PSTN Corpus are used for training and development. The rest 20% of Tencent Corpus and 5% of PSTN Corpus are used for evaluation test in this challenge. We aim

to make the impairment situation and score distribution in the divided dataset as even as possible. Meanwhile, as the NISQA corpus is already publicly available so they will only be used as part of the training and development test sets in this challenge. In addition, we will generate two evaluation test datasets (each with 200 clips, one in German and one in English) and subjective tests will be conducted using P.808 Toolkit.

In summary, there are about 86,000 speech clips for training and development, and 7,200 clips for evaluation test in this challenge. They are composed of Chinese, English, and German, and consider background noise, speech enhancement system, reverberation, codecs, packet-loss and other possible online conference voice impairment scenarios.

4. Challenge Rules and Requirements

4.1. Registration

The registration link is available on the challenge website. We kindly request participants to use institutional email for registration. Otherwise, the registration may be invalid. Once the registration is confirmed, participants will receive the confirmation letter of registration, and the information about downloading the challenge datasets.

Please note that any deliberate attempts to bypass the submission limit will lead to automatic disqualification. This includes, for example, creating multiple teams and submit multiple results by one participating team. In case of any issue, the final interpretation right belongs to the organizing committee.

4.2. Submission

4.2.1. Submitting the Predicted Score of Evaluation Test Set

Participants are required to submit the predicted score of evaluation test set. The filename corresponding to the predicted score should be kept the same as the provided evaluation test set. The rule details and way of submission will be notified by organizers in March 2022.

4.2.2. Submission of System Description

Each registered team is required to submit a technical system description report. Please prepare this report using the Interspeech 2022 paper template. Reports must be written in English. The system description does not need to repeat the content of the evaluation plan, such as the introduction of database, evaluation metric, etc. The system description should include the following items:

- a complete description of the system components, including the acoustic feature parameters, algorithm modules along with their configurations, etc.
- a complete data description for training. If extra training data besides those provided by the Challenge were used then detailed information of extra data and performance improvement after using extra data should be provided.
- the objective scores of clips in development and evaluation test set, including Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (SRCC) and RMSE. This challenge uses RMSE as the primary evaluation indicator.
- a report of the model size, real time factors (single threaded CPU, preferably Intel Core i5 quad core machine clocked at 2.4 GHz) as well as the amount of memory used to process a single clip.

4.2.3. Paper Submission

All participating teams should submit their papers to Interspeech 2022 special session - Non-intrusive Speech Quality Assessment in Online Conferencing Applications. (ConferencingSpeech 2022). Only the teams with papers submitted to ConferencingSpeech special session will be considered for the final ranking of this challenge. Please submit your paper by 21 Mar 2022 to the Interspeech 2022 paper submission system and choose this special session (ConferencingSpeech 2022). The papers will undergo the standard peer-review process of Interspeech 2022.

4.3. Other Important Rules

- Participants must abide by the rules in Section 2.
- There are no restrictions on the algorithm. Participants could develop any algorithm for the tasks.
- Participants must send the results achieved by their developed models to the organizers. The details of the submission can be found in Section 4.2. The final results on the evaluation test set must be included in the paper submitted to Interspeech 2022.
- Participants are forbidden to use any of the evaluation test set to fine-tune or retrain their models. Failing to adhere to these rules will lead to disqualification from the challenge.

5. Timeline

- Challenge registration open: January 19, 2022
- Release of evaluation plan, the list of training data and development test set: January 27, 2022
- Release of baseline system: February 17, 2022
- Deadline of challenge registration: March 12, 2022
- Release of evaluation test set: March 14, 2022
- Deadline of submitting the results: March 17, 2022
- Notification of the results of participants: March 20, 2022
- Interspeech paper submission deadline: March 21, 2022

6. Conclusion

7. References

[1] "https://www.magicdatatech.cn/datasets?"

[2] "http://www.openslr.org/38/."

[3] *http://www.aishelltech.com/kysjcp*.

[4] "https://github.com/gabrielmittag/nisqa/wiki/nisqa-corpus."

[5] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "Cosine-a corpus of multi-party conversational speech in noisy environments," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4153–4156.

[6] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout *et al.*, "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[7] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[8] X. Dong and D. S. Williamson, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," *INTERSPEECH*, pp. 4631–4635, 2020.

[9] C. K. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework."

[10] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[11] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[12] G. Mittag, R. Cutler, Y. Hosseinkashi, M. Revow, S. Srinivasan, N. Chande, and R. Aichner, "Dnn no-reference pstn speech quality prediction," *arXiv preprint arXiv:2007.14598*, 2020.