Interspeech 2022
18-22 September 2022, Incheon, Korea

ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications

Gaoxiong Yi[1], *Wei Xiao*[1], *Yiming Xiao*[1], *Babak Naderi*[2], *Sebastian Möller*[2], *Wafaa Wardah*[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

{gaoxiongyi,denniswxiao,yimingxiao,simeonshang}@tencent.com[1],{babak.naderi,sebastian.moeller,wafaa.wardah}@tu-berlin.de[2],{Ross.Cutler, gmittag}@microsoft.com[3],zhuozhan@iu.edu[4], williads@indiana.edu[4], fchen@sustech.edu.cn[5],fzhyang@mail.xidian.edu.cn[6]

1. Abstract

With the advances in speech communication systems such as online conferencing applications, we can seamlessly work with people regardless of where they are. However, during online meetings, speech quality can be significantly affected by background noise, reverberation, packet loss, and network jitter, to name a few. Because of its nature, speech quality is traditionally assessed in subjective tests in laboratories and lately also through crowdsourcing following the international standards from the ITU-T Rec. P.800 series. However, those approaches are costly and cannot be applied to customer data. Therefore, an effective objective assessment approach is needed to evaluate or monitor the speech quality of the ongoing conversation. The ConferencingSpeech 2022 challenge targets the non-intrusive deep neural network models for the speech quality assessment task. We open-sourced a training corpus with more than 86K speech clips in different languages, with a wide range of synthesized and live degradations and their corresponding subjective quality scores through crowdsourcing. 18 teams submitted their models for evaluation in this challenge. The blind test sets included about 4300 clips from wide ranges of degradations. This paper describes the challenge, the datasets, and the evaluation methods and reports the final results.

Index Terms: speech quality, deep learning, non-intrusive model

2. Introduction

With the popularity of remote conferencing, voice-based human-computer interaction has become mainstream. Environmental noise, room reverberation, digital signal processing, and network transmission can all degrade the quality of the speech signal. In these applications, speech quality assessment is in high demand. So far, the above fields have made great progress. In ITU-T Rec. P.800 [1], the international telecommunication union develops subjective evaluation procedures to assess speech quality, which is also the most preferred approach for quality assessment. However, it must be performed under controlled conditions, which is often time-consuming and expensive. Meanwhile, the perceptual evaluation of speech quality (PESQ) [2] and perceptual objective listening quality analysis (POLQA) [3] are designed to objectively evaluate speech quality. However, they need clean reference speech signals as comparison input. In order to non-intrusively assess the speech quality, ITU-T Rec. P.563 [4] was developed but only for target narrow-band applications. As deep learning shines in various fields, deep neural networks have been developed to address the non-intrusive speech quality assessment problem recently [5, 6, 7, 8, 9, 10, 11, 12]. Nevertheless, most of these methods adopt PESQ or POLQA as the speech quality label, which can not really represent the subjective ratings in all impairments. Only a few datasets with subjective scores have been published, which limits the application of deep learning in the above problem. Therefore, a large dataset with subjective speech quality scores and a non-intrusive speech quality assessment method, which can better reflect perceived subjective feelings, are urgently needed.

The ConferencingSpeech 2022 challenge aims to stimulate research in the above-mentioned areas. We provided comprehensive training and test datasets that contain at least 200 hours of speech samples with subjective test scores. We hope this challenge helps facilitate idea exchanges and discussions in this special session. Meanwhile, this challenge has the following features: 1) We aim for non-intrusive models for evaluating the speech quality (i.e., without reference speech signals), which is more practical in online conferencing applications. 2) 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. 3) To truly reflect subjective opinion on speech quality, the training and test datasets contain the mean opinion score (MOS), which is obtained through subjective absolute category rating tests via crowdsourcing and in accordance with the ITU-T Rec P.808 [13] using its open-sourced implementation [14]. 4) Different from the Clarity Prediction Challenge [15] which evaluates the speech intelligibility of speech signals, this is the first challenge on non-intrusive objective speech quality assessment in an online conferencing. We provide speech clips with subjective MOS that covers most of the impairment scenarios in on-line speech communication. It is believed that this will promote the development of non-intrusive objective speech quality assessment methods.

10.21437/Interspeech.2022-10597

Table 1: *Proportion of the degradations applied in Tencent Corpus.*

Impairment	Percentage
White noise	10%
Nonstationary background noise	60%
High-pass/low-pass filtering	3.75%
Amplitude clipping	1.25%
AMR [16]/Opus [17] 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: *Proportion of the second step simulated impaired speech in Tencent Corpus.*

Impairment	Percentage
Only first step impairments	16%
First step + noise suppression	49.2%
First step + noise suppression + packet loss concealment	23.9%
Clean speech	4.8%
Clean speech + packet loss concealment	6.1%

3. Task Description

In this challenge, comprehensive training datasets with ground truth MOS were provided to each registered team. It is anticipated that the participating teams use only the impaired speech signals 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) and Pearson correlation coefficient (PCC). More details can be found on the challenge website [1].

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 should be included that shows the benefit to the test set.

4. 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 the NISQA Corpus, the other three datasets are all made public for the first time. Each dataset will be described in detail below.

4.1. Tencent Corpus

This dataset includes speech conditions with reverberation and without reverberation. In the without reverberation condition, there are about 10000 Chinese speech signals with simulated impairments, which is very common in an online conference. In the with reverberation condition, a total of approximately 4000 simulated impairments and live recording speech clips are both considered. Part of the Tencent corpus speech samples are recorded at 16KHz, while the remaining are recorded at 48KHz.

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. In order to prevent the possible speaker-dependent behavior of the trained model, the original speech data was selected from three publicly available datasets Magic data [18], ST Mandarin [19] and AIshell_100h [20]. Each speech clip in the source data was processed with one type of impairment and only one type. The different impairment types and the corresponding percentage of the speech clips applied with each impairment type are listed in Table 1. Based on the speech clips processed in the above step, we applied another speech processing step including noise suppression [21] and packet loss concealment [22] to simulate more realistic online communication. Those processing and corresponding percentage in the final dataset are listed in Table 2.

In order to make the subjective database more comprehensive, 4000 speech clips with reverberation were added to the dataset. 28% of them were generated with simulated reverberation and 72% were recorded in realistic reverberant rooms. In the simulated reverberation condition, the source data came from the purchased king-asr-166 dataset. Meanwhile, various room sizes and reverberation delays were considered. The subjective scoring procedure was conducted in a crowdsourcing way similar to ITU-T P.808. 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.

4.2. NISQA Corpus

The NISQA Corpus includes more than 14000 speech samples. Part of the NISQA corpus speech samples are recorded at 16KHz while the remaining are recorded at 48KHz sampling rates. The corpus is already publicly available therefore we only included it in the training and development test sets in the competition. Subjective ratings were collected through an extension of the P.808 Toolkit [14]. Each clip has on average 5 valid votes. Further details about this corpus are provided in [23]. We also created a new test dataset (TUB hereafter) using unimpaired signals of 136 conversation tests. We selected a portion of speech with no overlaps between two speakers, at least 55% active speech, and added leading and trailing silences. That led to 865 source clips, from which a basic clustering algorithm detected seven different clusters. Meanwhile, we created 62 synthetic degradation conditions (different codecs, bandwidths, single or multiple background noises, packet lost scenarios, etc.). Each condition was applied on seven randomly selected source clips (one per cluster). Finally, we collected on average 18 subjective ratings per clip using the P.808 Framework [14].

4.3. IU Bloomington Corpus

There are 36000 speech clips extracted from COSINE [24] and VOiCES [25] datasets. We randomly select about 10000 clips form the IU Bloomington Corpus in this challenge. For the 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 are between -4.9 to 4.3 dB. Three versions of each speech utterance were provided for the 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. We crowdsourced our listening tests on Amazon Mechanical Turk by publishing 700 human intelli-

[1] https://tea-lab.qq.com/conferencingspeech-2022

gence tasks following ITU-R BS.1534 [26]. The speech corpora from IU Bloomington Corpus consist of 16-bit single channel files sampled at 16 kHz. For more details, please refer to [27].

4.4. PSTN Corpus

The clean reference files used for the phone calls are derived from the public audiobook dataset Librivox. Because many of the recordings are of poorer quality, the files have been filtered according to their quality as described in [28], leaving in a total 441 hours from 2150 speakers of good quality speech. We randomly select about 100 hours clips form the PSTN Corpus in this challenge. Since, in practice, there are often environmental sounds present during phone calls, we used the DNS Challenge 2021 [28] to add background noise. The noise clips are taken from Audioset [29], Freesound, and the DEMAND [30] corpus and added to the clean files with an SNR between $0 - 40$ dB. The perceived speech quality of the training and test sets were annotated in a listening experiment on AMT, according to P.808. The speech corpora from PSTN Corpus are sampled at 8 kHz. For more details, please refer to [31].

4.5. Dataset Division

The training, development, and evaluation test sets in this challenge are all originated from the above-mentioned datasets. It is worth noting that the IU Bloomington corpus differs from the Tencent, NISQA and PSTN corpora that used ITU-T P.808 for subjective testing, where the IU Bloomington corpus adopted ITU-R BS.1534 for subjective testing, which resulted in a rating range of 0$\sim$100 instead of 1$\sim$5. 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, 5% of PSTN Corpus, and newly created TUB 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. In summary, there are about 86000 speech clips about 191 hours for training and development, and 4372 clips about 11 hours for the 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. More details about the dataset in this challenge can be found in the Challenge Evaluation Plan[2]

5. Baseline System

This challenge provided two baseline systems, including Baseline System 1 and Baseline System 2 [23]. The Baseline System 1 is a simplified version of the model in [32]. It is made of a deep feed forward network followed by long short-term memory and average pooling. The Baseline System 2 is the complete model from [23]. It is based on a convolutional neural network (CNN) and attention mechanism. The log-mel-spectrograms of the speech signal is provided as input to the CNN network to extract the speech quality features at different times. The estimated per-frame quality values are then aggregated over time by using an attention model and the final score is predicted by the attention pooling block. Both Baseline Systems were trained on all datasets provided in this challenge.

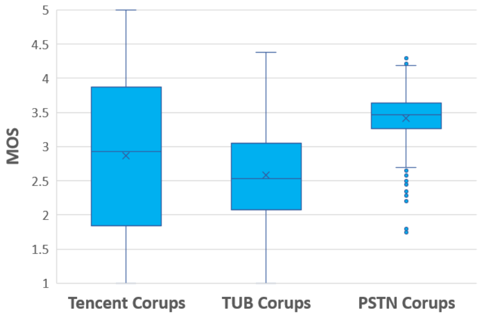


Figure 1: *Distribution of MOS values in three blind test sets.*

6. Challenge Results

6.1. Evaluation Setup and Results

According to ITU-T P.1401 [33], we calculated RMSE to evaluate the accuracy, the outlier ratio (OR) for consistency and PCC for linearity. The RMSE is calculated according to the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} Perror(i)^2}, \tag{1}$$

where N denotes the total number of speech utterances, i indicates the i-th speech signal. $Perror(\cdot)$ represents the prediction error which is defined as the difference between the measured and predicted MOS:

$$Perror(i) = \mathrm{MOS}(i) - \mathrm{MOS}_p(i). \tag{2}$$

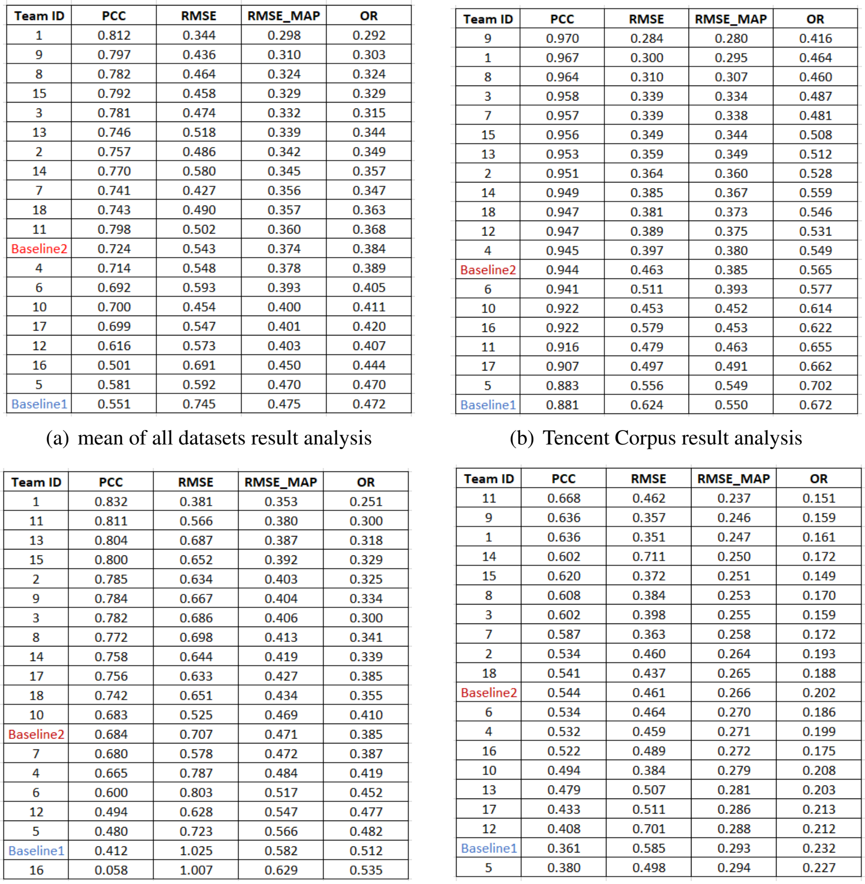
The OR represents the number of outlier-points to the total number of speech utterances. An outlier is defined as a point for which the prediction error is larger than the 95% confidence interval of the MOS value. Meanwhile, due to bias or offsets, different gradients and different qualitative rank orders are always present in subjective evaluations. The statistical uncertainty always exists in the collected MOS. Therefore, a mapping function is recommended to compensate for the possible variance between several subjective experiments. In this challenge, a third-order polynomial function is applied, which can be derived from $y'' = a+by+cy^2+dy^3$, where rmse $(x, y'') \rightarrow \min$ and $f(y) =$ monotonous between $y''_{\min}$ and $y''_{\max}$.

For each model, we create one mapping function per test dataset. Then the mapped predictions were used to calculated RMSE_MAP and OR. In this challenge, we decided to rank models based on their accuracy, i.e., RMSE_MAP. The descriptive statistics on subjective ratings per blind test set are provided in Table 3 and Figure 1. A total of 18 teams from different countries submitted results. Based on the submitted results, their PCC, RMSE, RMSE_MAP, and OR are calculated. The specific results are shown in Figure 2.

6.2. Key takeaways

- In Figure 2 (a), it can be observed that the prediction results of all teams are better than the results of the Baseline System 1. It suggests that the generalization of Baseline System 1 is not good enough to effectively cover

[2] https://tea-lab.qq.com/ConferencingSpeech_2022_Challenge_Evaluation_Plan_version2.pdf



Team ID	PCC	RMSE	RMSE_MAP	OR
1	0.812	0.344	0.298	0.292
9	0.797	0.436	0.310	0.303
8	0.782	0.464	0.324	0.324
15	0.792	0.458	0.329	0.329
3	0.781	0.474	0.332	0.315
13	0.746	0.518	0.339	0.344
2	0.757	0.486	0.342	0.349
14	0.770	0.580	0.345	0.357
7	0.741	0.427	0.356	0.347
18	0.743	0.490	0.357	0.363
11	0.798	0.502	0.360	0.368
Baseline2	0.724	0.543	0.374	0.384
4	0.714	0.548	0.378	0.389
6	0.692	0.593	0.393	0.405
10	0.700	0.454	0.400	0.411
17	0.699	0.547	0.401	0.420
12	0.616	0.573	0.403	0.407
16	0.501	0.691	0.450	0.444
5	0.581	0.592	0.470	0.470
Baseline1	0.551	0.745	0.475	0.472

(a) mean of all datasets result analysis

Team ID	PCC	RMSE	RMSE_MAP	OR
9	0.970	0.284	0.280	0.416
1	0.967	0.300	0.295	0.464
8	0.964	0.310	0.307	0.460
3	0.958	0.339	0.334	0.487
7	0.957	0.339	0.338	0.481
15	0.956	0.349	0.344	0.508
13	0.953	0.359	0.349	0.512
2	0.951	0.364	0.360	0.528
14	0.949	0.385	0.367	0.559
18	0.947	0.381	0.373	0.546
12	0.947	0.389	0.375	0.531
4	0.945	0.397	0.380	0.549
Baseline2	0.944	0.463	0.385	0.565
6	0.941	0.511	0.393	0.577
10	0.922	0.453	0.452	0.614
16	0.922	0.579	0.453	0.622
11	0.916	0.479	0.463	0.655
17	0.907	0.497	0.491	0.662
5	0.883	0.556	0.549	0.702
Baseline1	0.881	0.624	0.550	0.672

(b) Tencent Corpus result analysis

Team ID	PCC	RMSE	RMSE_MAP	OR
1	0.832	0.381	0.353	0.251
11	0.811	0.566	0.380	0.300
13	0.804	0.687	0.387	0.318
15	0.800	0.652	0.392	0.329
2	0.785	0.634	0.403	0.325
9	0.784	0.667	0.404	0.334
3	0.782	0.686	0.406	0.300
8	0.772	0.698	0.413	0.341
14	0.758	0.644	0.419	0.339
17	0.756	0.633	0.427	0.385
18	0.742	0.651	0.434	0.355
10	0.683	0.525	0.469	0.410
Baseline2	0.684	0.707	0.471	0.385
7	0.680	0.578	0.472	0.387
4	0.665	0.787	0.484	0.419
6	0.600	0.803	0.517	0.452
12	0.494	0.628	0.547	0.477
5	0.480	0.723	0.566	0.482
Baseline1	0.412	1.025	0.582	0.512
16	0.058	1.007	0.629	0.535

(c) TUB Corpus result analysis

Team ID	PCC	RMSE	RMSE_MAP	OR
11	0.668	0.462	0.237	0.151
9	0.636	0.357	0.246	0.159
1	0.636	0.351	0.247	0.161
14	0.602	0.711	0.250	0.172
15	0.620	0.372	0.251	0.149
8	0.608	0.384	0.253	0.170
3	0.602	0.398	0.255	0.159
7	0.587	0.363	0.258	0.172
2	0.534	0.460	0.264	0.193
18	0.541	0.437	0.265	0.188
Baseline2	0.544	0.461	0.266	0.202
6	0.534	0.464	0.270	0.186
4	0.532	0.459	0.271	0.199
16	0.522	0.489	0.272	0.175
10	0.494	0.384	0.279	0.208
13	0.479	0.507	0.281	0.203
17	0.433	0.511	0.286	0.213
12	0.408	0.701	0.288	0.212
Baseline1	0.361	0.585	0.293	0.232
5	0.380	0.498	0.294	0.227

(d) PSTN Corpus result analysis

Figure 2: *Challenge result analysis*

Table 3: *Descriptive statistics on subjective ratings in blind test sets.*

Dataset	Average No. ratings p. clip	Average 95%CI	MOS min	MOS max
Tencent Corpus	28	0.20	1.00	5.00
TUB Corpus	18	0.40	1.00	4.37
PSTN Corpus	24	0.35	1.74	4.29

different datasets. Meanwhile 11 teams achieved better results than the Baseline System 2, accounting for 61% of the teams who submitted their results.

- The rank-order of models change based on the dataset and the criteria to use. For any future benchmarking, we recommend to consider multiple blind test sets created by different laboratories. PCC is strongly dependent to how well the MOS values in the test set are distributed: For Tencent Corpus team reach a PCC in range of [0.88 , 0.97] whereas in the other two datasets the achieved PCC was strongly smaller (cf. Figure 1). The OR is directly influenced by the 95% CI, therefore OR values from one dataset cannot be directly compare to another dataset.
- We did not observe a significant difference between mapped RMSE values of top 2, 8, and 9 models in datasets Tencent, TUB, and PSTN, respectively. However, team #1 is consistently among the top-three for all datasets.
- From the top-performing team methods, it can be observed that a large pre-trained language model is usually used for feature extraction, followed by a downstream network to fit the MOS score.

7. Conclusion

The ConferencingSpeech 2022 Challenge was organized to help researchers from academia and industry to facilitate the development of non-intrusive objective speech quality assessment for online conferencing applications. We open-sourced several large training datasets with subjective scores. We recommend considering different blind test sets, created by multiple groups, for any similar challenges or benchmarking tasks.

8. Acknowledgement

The work performed by authors at Technical University of Berlin was supported by the German Research Foundation (DFG) within project MO 1038/32-1.

9. References

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

[2] A. Rix, J. Beerends, M. Hollier, and A. 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.*, vol. 2, 2001, pp. 749–752 vol.2.

[3] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[4] L. Malfait, J. Berger, and M. Kastner, "P.563—the ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[5] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 2315–2319.

[6] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, 2018.

[7] S. Fu, Y. Tsao, H. Hwang, and H. Wang, "Quality-Net: an end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech 2018*, 2018.

[8] A. H. Andersen, J. M. De Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, 2018.

[9] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019.

[10] H. Gamper, C. K. 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)*. IEEE, 2019, pp. 85–89.

[11] 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)*. IEEE, 2019, pp. 631–635.

[12] X. Dong and D. S. Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 100–104.

[13] ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*. Geneva: International Telecommunication Union, 2021.

[14] B. Naderi and R. Cutler, "An Open Source Implementation of ITU-T Recommendation P.808 with Validation," in *Proc. Interspeech 2020*, 2020.

[15] "https://claritychallenge.github.io/clarity/cpc1/doc/."

[16] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE transactions on speech and audio processing*, vol. 10, no. 8, pp. 620–636, 2002.

[17] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," *IETF, September*, vol. 2, 2012.

[18] "Magic Data," https://www.magicdatatech.cn/datasets, accessed: 2022-02-25.

[19] "SLR38: Free ST Chinese Mandarin Corpus," http://www.openslr.org/38/, accessed: 2022-02-25.

[20] "AISHELL- Open Source Mandarin Speech Corpus," http://www.aishelltech.com/kysjcp, accessed: 2022-02-25.

[21] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[22] ITU-T Recommendation G.191, *Software Tools for Speech and Audio Coding Standardization*. Geneva: International Telecommunication Union, 2019.

[23] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. Interspeech 2021*, 2021.

[24] 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.

[25] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (VOiCES) corpus," in *Proc. Interspeech 2018*, 2018.

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

[27] X. Dong and D. S. Williamson, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in *Proc. INTERSPEECH*, 2020, pp. 4631–4635.

[28] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech 2020*, 2020.

[29] 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.

[30] 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.

[31] G. Mittag, R. Cutler, Y. Hosseinkashi, M. Revow, S. Srinivasan, N. Chande, and R. Aichner, "DNN no-reference PSTN speech quality prediction," in *Proc. Interspeech 2020*, 2020.

[32] 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)*. IEEE, 2019, pp. 7125–7129.

[33] ITU-T Recommendation P.1401, *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. Geneva: International Telecommunication Union, 2020.