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

NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets

Gabriel Mittag[1], *Babak Naderi*[1], *Assmaa Chehadi*[1], *Sebastian Möller*[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
first.last@tu-berlin.de

Abstract

In this paper, we present an update to the NISQA speech quality prediction model that is focused on distortions that occur in communication networks. In contrast to the previous version, the model is trained end-to-end and the time-dependency modelling and time-pooling is achieved through a Self-Attention mechanism. Besides overall speech quality, the model also predicts the four speech quality dimensions *Noisiness*, *Coloration*, *Discontinuity*, and *Loudness*, and in this way gives more insight into the cause of a quality degradation. Furthermore, new datasets with over 13,000 speech files were created for training and validation of the model. The model was finally tested on a new, live-talking test dataset that contains recordings of real telephone calls. Overall, NISQA was trained and evaluated on 81 datasets from different sources and showed to provide reliable predictions also for unknown speech samples. The code, model weights, and datasets are open-sourced.

Index Terms: speech quality, deep learning

1. Introduction

One of the main performance indicators for the evaluation of telecommunication networks is the perceived speech quality. It is traditionally derived from subjective listening tests according to ITU-T P.800 [1] or recently also through crowdsourced listening tests according to ITU-T P.808 [2,3]. The average rating across all test participants for a speech sample then gives the mean opinion score (MOS). However, because listening tests are costly and time consuming, instrumental models have been developed that can predict the speech quality automatically. The currently recommended model by the ITU-T is POLQA [4], which requires the clean reference and the degraded output signal to predict the speech quality based on a comparison of both signals. In contrast to these types of *double-ended* models, *single-ended* models only require the degraded output signal, which makes it possible to monitor the quality of live phone calls, or predict the speech quality of samples for which no clean reference is available. However, the currently recommended single-ended speech quality by the ITU-T, P.563 [5], is only available for narrowband (NB) speech signals. Furthermore, the prediction performance showed to decrease for conversational speech and modern VoIP (Voice over IP) distortions that were not present when P.563 was developed [6].

While the overall MOS is an important indicator, it gives no insight into the cause of a quality degradation. To overcome this problem, it was shown in [7] that the speech quality multidimensional space of modern communication networks is made up of the three orthogonal dimensions: *Noisiness*, *Coloration*, and *Discontinuity*. Later, the *Loudness* was added as a fourth dimension in [8], although it is not entirely orthogonal to the other dimensions. These four perceptual dimensions can be quantified through auditory listening tests and are linked to technical root causes.

Recently, deep learning methods have been applied to build single-ended speech quality models [9–17] and showed to outperform traditional approaches without the need for a clean reference. In [18], we presented the deep learning model NISQA that predicts speech quality of super-wideband (SWB) speech samples. In [19], we presented an extension of the model that predicts three of the four quality dimensions based on expert scores due to the lack of available subjective data. In this paper, we present an update to NISQA that predicts the overall MOS and the four speech quality dimensions with one multi-task neural network. Furthermore, we created a large pool of eight new speech quality datasets with subjective MOS and quality dimension ratings for training and evaluation. Because of the increased available data, we could train the model end-to-end with subjective data only, without the need for objective MOS values. Also, we improved the neural network architecture of the model by replacing the CNN-LSTM structure with a CNN–Self-Attention–Attention-Pooling (CNN-SA-AP) network. The model is overall trained and evaluated on a large set of 81 datasets from different sources. Another advantage of the updated NISQA model is that it can be applied to speech samples of any duration or sample rate without any preprocessing steps or level normalisation. Finally, the PyTorch code, the model weights, and several of the datasets are open-sourced on GitHub[1].

2. Method

The model can be divided into four stages: 1) Mel-Spec segmentation, 2) Framewise model (CNN), 3) Time-Dependency model (Self-Attention), 4) Pooling model (Attention-Pooling). An overview of this architecture is shown in Figure 1. In the first stage, Mel-specs are calculated from the input signal and then divided into overlapping segments. In the second stage, a framewise neural network with the Mel-spec segments as inputs is used to compute features that are suitable for speech quality prediction. These features are calculated on a frame basis and therefore result in a sequence of framewise features. In a third stage, the time dependencies of the feature sequence are modelled. Finally, the features are aggregated over time in a pooling layer. The aggregated features are then used to estimate a single MOS value.

In this paper, different neural network architectures are applied and compared for each model stage. In the ablation study (Sect. 4), we show that the best combination is a CNN as a framewise model, a Self-Attention network as a time-dependency model, and Attention-Pooling as a pooling model.

[1] www.github.com/gabrielmittag/NISQA

http://dx.doi.org/10.21437/Interspeech.2021-299

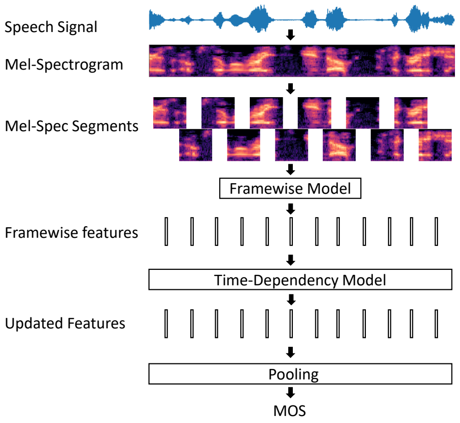

Figure 1: *General speech quality model structure.*

More details about the neural network structure can be found in the open-sourced code.

2.1. Mel-Spec Segmentation

The input to the model are Mel-spec segments with 48 Mel-bands. The FFT window length is 20 ms with a hop size of 10 ms. The maximum frequency was chosen to be 20 kHz to be able to predict speech quality for up to fullband (FB). The Mel-specs are divided into segments with a width of 15 (i.e. 150 ms) and a height of 48. The hop size between the segments is 4 (40 ms), which leads to a segment overlap of 73% and overall 250 segments for a 10-second speech signal. The framewise network is provided with this wider segment of 150 ms to give the network some contextual awareness. The short-term and long-term temporal modelling, however, follows in the third model stage.

2.2. Framewise Model

As framewise model the CNN from [18] is used. It contains 6 convolutional layers and 3 max-pooling layers that downsample the input of dimension 48 × 15 to a size of 6 × 3. The final convolutional layer does not apply a width padding and therefore further reduces the output dimension to 64 × 6 × 1, where 64 represent the number of kernels. Finally, the output is flattened. Thus, each Mel-spec segment results in a feature vector of length 384 after passing the CNN. As a baseline model, a basic deep feedforward network with a depth of four layers and 2048 hidden units each is implemented.

2.3. Time Dependency

In this stage, the individual time steps of the feature sequence can interact with each other to improve the prediction performance. To this end, a Self-Attention network is applied, which is based on the Transformer encoder [20]. Because the Self-Attention only models temporal dependencies of framewise features that are already computed by a deep framewise network, a relatively low complexity of the Transformer is sufficient. Moreover, in practice, it was noted that the multi-head mechanism did not improve the results. Therefore, the block is implemented with a single head, a depth of 2 blocks, a model dimensionality of $d_{\mathrm{tf}} = 64$, and a feedforward network with $d_{\mathrm{tf,ff}} = 64$ hidden units. As a baseline, a single BiLSTM layer with 128 hidden units in each direction is used.

2.4. Pooling

The quality of a telephone call can generally not be predicted accurately by simply taking the average quality across time. As was shown in [21], the “recency effect” and the out-weighting of poor quality segments in a call have to be considered adequately. Therefore, we propose to use an attention mechanism for time pooling. An overview of the attention-pooling block is shown in Figure 2, where y is the output $d_{\mathrm{tf}}\mathrm{x}L$-matrix of the time-dependency model. The matrix contains a zero-padded sequence of length L with feature vectors of dimension d_{tf}. The feedforward network with an output size of 1 and 128 hidden units is applied to each time step separately and identically. The attention scores computed by the feedforward network are then masked at the zero-padded time steps and applied to a softmax function to yield the normalised attention weights. These weights are applied to the input matrix y with a matrix multiplication operation. The weighted average feature vector z is then finally passed through a fully connected layer to estimate the overall speech quality. As baseline models, *average-pooling* and *max-pooling* are applied.

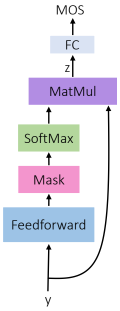

Figure 2: *Attention-pooling block*

2.5. Multidimensional model

The multidimensional prediction can be seen as a Multi-Task-Learning problem. The pooling block is computed separately for each dimension and overall MOS, while the CNN and Self-Attention network is shared across all tasks. Figure 3 shows how the Mel-spec features are calculated by the same CNN and Self-Attention network for each dimension. The outputs of each Self-Attention time-step are then the input for five individual pooling blocks that predict the overall MOS and the dimension scores.

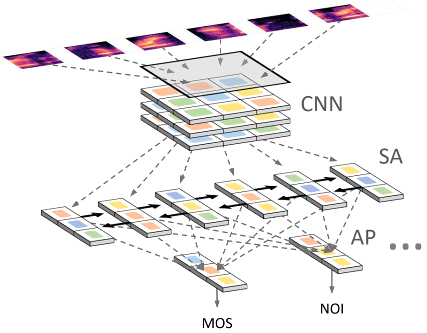

Figure 3: *NISQA neural network architecture.*

3. Datasets

NISQA is trained and evaluated on a large set of 59 training datasets (72,903 files), 18 validation sets (9,567 files), and 4 test sets (952 files)[2]. 55 of the datasets are taken from the POLQA Pool [4], 7 of the datasets are taken from the ITU-T P Suppl. 23 [22] pool, 11 datasets are older, internal speech quality datasets. Additionally, for this work 8 new datasets with overall quality and quality dimension ratings, and a large variety of different speakers, were created. 2 training datasets NISQA_TRAIN_SIM (10,000 samples from 2,322 speakers), NISQA_TRAIN_LIVE (1,020 samples from 486 speakers) and 2 validation datasets NISQA_VAL_SIM (2,500 samples from 938 speakers), NISQA_VAL_LIVE (200 samples from 102 speakers). The clean source speech samples are taken from four different English speech corpora: AusTalk [23] (containing conversational speech taken from interviews), DNS-Challenge [24] (LibriVox audiobooks), TSP [25] (read out Havard sentences [26]), and UK-Ireland dataset [27] (read out public domain texts). The datasets NISQA_TRAIN_SIM and NISQA_VAL_SIM contain simulated speech distortions, such as packet-loss, bandpass filter, different codecs, and clipping. To simulate real background noises, the noise clips from the DNS-Challenge datasets were used, which in turn are taken from the Audioset [28], freesound [29], and DEMAND [30] corpora. The datasets NISQA_TRAIN_LIVE and NISQA_VAL_LIVE contain live Skype and landline-to-mobile phone recordings, where the LibriVox audiobook reference files were played back through a loudspeaker directly into the terminal device (phone/laptop). During the recording different real distortions were created, such as typing on a keyboard, opening window (street noise), or poor reception. These four training and validation datasets were annotated in the crowd according to ITU-T P.808 [2] with 5 ratings per file.

Additionally, four independent test sets were created that were not considered before the final training of the NISQA model. NISQA_TEST_P501, NISQA_TEST_FOR, NISQA_TEST_NSC contain simulated distortions and additionally live VoIP calls with Zoom, Skype, Google Meet, WhatsApp, and Discord, where the reference speech samples were played back directly from the laptop. Then a poor internet connection was simulated to obtain files with different distortions (packet-loss, warping, low-bitrate). NISQA_TEST_P501 contains the English Annex C files from ITU-T P.501 [31], NISQA_TEST_FOR contains English conversation samples from the Forensic Voice dataset [32, 33], NISQA_TEST_NSC contains German conversational speech from the NSC dataset [34]. These three datasets were again annotated in the crowd according to ITU-T P.808 with 30 ratings per file.

Because the use-case for a single-ended prediction model are real phone calls with conversational speech, from an unknown speaker, device, and network, a fourth dataset NISQA_TEST_LIVETALK with real phone call recordings was created, where the talkers spoke directly into the terminal device (i.e. a smartphone or laptop). The test participants were instructed to talk loudly, quietly, with loudspeaker, or music in the background to obtain different test scenarios and speech quality distortions. Depending on the condition the talkers were located in different environments, such as in a café, inside a car on the highway, inside a building with poor reception, elevator, shopping centre, subway/metro station, on a busy street, etc. The talkers used their mobile phone to call either through the mobile network or with a VoIP service (Skype/Facebook). The dataset consists of 58 different conditions with each 4 different files, resulting in 232 files overall. The speech files were recorded from 8 different talkers (4 male and 4 female) in German, where for each condition 2 male and 2 female talkers were selected. The dataset was annotated in the lab according to P.800 with 24 ratings per file.

4. Ablation Study

In this section, it will be shown that a combination of a CNN as a framewise model, a Self-Attention network as time-dependency model, and an Attention-Pooling network as pooling model (CNN-SA-AP) gives the highest prediction performance. To this end, an ablation study is performed, in which one of the three neural network model stages is either removed or replaced with another network type. For training, the simulated and live training sets NISQA_TRAIN_SIM and NISQA_TRAIN_LIVE were used. The models are evaluated on the average PCC (Pearson's correlation coefficient) between the validation datasets NISQA_VAL_SIM and NISQA_VAL_LIVE. The training for each model configuration is run 12 times to rule out random effects, the median performance over these 12 training runs is presented in the following result tables.

4.1. Framewise Model

Table 1 shows the results when different framewise models are applied. The *CNN*-SA-AP model clearly outperforms the basic feedforward neural network (*FFN*-SA-AP) and the model without framewise network that only applies Self-Attention and Attention-Pooling (*Skip*-SA-AP).

Table 1: *Framewise model comparison with Median PCC.*

Model	Skip	CNN	FFN
r	0.772	**0.870**	0.812

4.2. Time-Dependency Model

Table 2 shows the results for different time-dependency models. Again, it can be seen that the CNN-*SA*-AP model achieves the best results. However, the difference between Self-Attention and CNN-*LSTM*-AP is only small. Further, it can be seen that a combination of SA and LSTM worsens the results, compared to SA or LSTM only. Although the difference between SA and LSTM is small, the overall performance can be notably increased when compared to the model CNN-*Skip*-AP without time-dependency modelling.

Table 2: *Time-Dependency comparison with Median PCC.*

Model	Skip	SA	LSTM	LSTM-SA	SA-LSTM
r	0.851	**0.870**	0.866	0.862	0.863

4.3. Pooling

Table 3: *Pooling comparison with Median PCC.*

Model	AP	Avg	Max
r	**0.870**	0.867	0.866

[2] A detailed overview of the individual datasets can be found on the GitHub.

Table 4: *Per-condition validation and test results of the overall quality in terms of PCC and RMSE after first-order mapping.*

Dataset	Scale	Lang	Con	Files	NISQA		P563		ANIQUE+		WAWEnets		POLQA		DIAL		VISQOL	
					r	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE
103_ERICSSON	SWB	se	54	648	0.85	0.38	0.36	0.66	0.54	0.60	0.28	0.68	**0.87**	**0.34**	0.78	0.45	0.26	0.69
104_ERICSSON	NB	se	55	660	0.77	0.47	0.64	0.57	0.68	0.55	0.13	0.74	**0.91**	**0.31**	0.76	0.49	0.39	0.69
203_FT_DT	SWB	fr	54	216	**0.92**	**0.36**	0.68	0.69	0.47	0.82	0.64	0.72	0.91	0.38	0.79	0.57	0.59	0.75
303_OPTICOM	SWB	en	54	216	0.92	0.33	0.85	0.44	0.71	0.59	0.43	0.76	**0.93**	**0.31**	0.71	0.59	0.42	0.76
403_PSYTECHNICS	SWB	en	48	1152	0.91	0.36	0.81	0.50	0.77	0.54	0.78	0.53	**0.96**	**0.24**	0.92	0.34	0.73	0.57
404_PSYTECHNICS	NB	en	48	1151	0.77	0.39	0.82	0.35	0.74	0.41	0.14	0.61	**0.86**	**0.31**	0.67	0.46	0.55	0.51
503_SWISSQUAL	SWB	de	54	216	0.92	0.34	0.71	0.62	0.61	0.70	0.59	0.71	**0.94**	**0.29**	0.85	0.46	0.65	0.67
504_SWISSQUAL	NB	de	49	196	**0.92**	**0.37**	0.83	0.50	0.79	0.56	0.54	0.77	0.87	0.45	0.73	0.63	0.60	0.73
603_TNO	SWB	nl	48	192	0.89	0.44	0.83	0.53	0.69	0.69	0.59	0.77	**0.95**	**0.29**	0.86	0.48	0.47	0.84
ERIC_FIELD_GSM_US	NB	en	372	372	**0.79**	**0.36**	0.42	0.54	0.17	0.58	0.60	0.47	0.75	0.39	0.71	0.42	0.51	0.51
HUAWEI_2	NB	zh	24	576	**0.98**	**0.21**	0.93	0.35	0.79	0.59	0.63	0.75	0.94	0.32	0.89	0.44	0.97	0.24
ITU_SUPPL23_EXP1o	NB	en	44	176	0.92	0.31	0.90	0.34	**0.98**	**0.15**	0.73	0.53	0.91	0.32	0.91	0.33	0.86	0.39
ITU_SUPPL23_EXP3d	NB	ja	50	200	0.92	0.27	0.93	0.26	**0.97**	**0.17**	0.68	0.50	0.85	0.36	0.84	0.36	0.79	0.41
ITU_SUPPL23_EXP3o	NB	en	50	200	0.91	0.30	0.91	0.30	**0.98**	**0.15**	0.79	0.45	0.88	0.35	0.87	0.36	0.78	0.45
TUB_AUS	FB	en	50	600	**0.91**	**0.21**	0.62	0.40	0.65	0.39	0.70	0.36	0.88	0.24	0.73	0.35	0.63	0.40
TUB_LIKE	SWB	de	8	96	0.98	0.25	0.85	0.60	0.85	0.61	0.59	0.93	**0.99**	**0.16**	0.89	0.53	0.81	0.67
NISQA_VAL_LIVE	FB	en	200	200	**0.82**	**0.40**	0.42	0.64	0.51	0.61	0.36	0.66	0.67	0.52	-0.22	0.69	0.66	0.53
NISQA_VAL_SIM	FB	en	2500	2500	**0.90**	**0.48**	0.45	0.99	0.54	0.93	0.30	1.05	0.86	0.56	0.36	1.03	0.78	0.69
NISQA_TEST_P501	FB	en	60	240	**0.95**	0.31	0.72	0.67	0.73	0.66	0.80	0.59	**0.95**	**0.30**	0.80	0.59	0.80	0.58
NISQA_TEST_NSC	FB	de	60	240	**0.97**	**0.23**	0.69	0.67	0.62	0.74	0.78	0.59	0.93	0.35	0.79	0.57	0.78	0.59
NISQA_TEST_FOR	FB	en	60	240	**0.95**	**0.26**	0.52	0.71	0.54	0.70	0.81	0.49	0.92	0.33	0.75	0.55	0.68	0.61
NISQA_TEST_LIVETALK	FB	de	58	232	**0.90**	**0.35**	0.70	0.58	0.56	0.68	0.66	0.61	N/A	N/A	N/A	N/A	N/A	N/A

Table 5: *Per-condition validation and test results of the speech quality dimensions in terms of PCC and RMSE after first-order mapping.*

Dataset	NOISINESS				COLORATION				DISCONTINUITY				LOUDNESS			
	NISQA		DIAL		NISQA		DIAL		NISQA		DIAL		NISQA		DIAL	
	r	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE
503_SWISSQUAL	**0.94**	**0.26**	0.84	0.39	0.84	0.39	**0.88**	**0.34**	**0.86**	**0.31**	0.74	0.42	0.91	0.29	**0.94**	**0.23**
TUB_AUS	**0.97**	**0.16**	0.88	0.29	**0.84**	**0.28**	0.81	0.3	**0.92**	**0.23**	0.61	0.46	**0.74**	**0.32**	0.62	0.38
NISQA_VAL_LIVE	**0.73**	**0.49**	0.31	0.69	**0.57**	**0.43**	-0.11	0.51	**0.55**	**0.56**	0.10	0.67	**0.73**	**0.47**	0.54	0.58
NISQA_VAL_SIM	**0.86**	**0.48**	0.40	0.87	**0.84**	**0.50**	0.25	0.90	**0.84**	**0.54**	0.23	0.97	**0.81**	**0.48**	0.38	0.76
NISQA_TEST_P501	**0.95**	**0.30**	0.86	0.48	**0.91**	**0.31**	0.62	0.60	**0.91**	**0.37**	0.68	0.65	**0.95**	**0.26**	0.88	0.40
NISQA_TEST_NSC	**0.96**	**0.22**	0.78	0.50	**0.93**	**0.28**	0.70	0.55	**0.96**	**0.30**	0.79	0.63	**0.96**	**0.25**	0.93	0.34
NISQA_TEST_FOR	**0.95**	**0.23**	0.70	0.50	**0.94**	**0.24**	0.67	0.53	**0.97**	**0.25**	0.81	0.55	**0.96**	**0.20**	0.87	0.36
NISQA_TEST_LIVETALK	**0.76**	**0.47**	N/A	N/A	**0.87**	**0.31**	N/A	N/A	**0.83**	**0.40**	N/A	N/A	**0.71**	**0.36**	N/A	N/A

The results for different pooling mechanism can be seen in Table 3, where a CNN as framewise and SA as time-dependency model is applied. The performance difference between the analysed pooling mechanism is only marginal, however, Attention-Pooling slightly performs better than Average- or Max-Pooling.

5. Results

The final model was trained on a set of 59 training and 18 validation datasets with a batch size of 160, learning rate of 0.001, Adam optimiser and bias-aware loss according to [35]. After each epoch, the model weights were stored, and the results on the training and validation set were calculated as average PCC across all datasets. The training was stopped after the validation PCC did not increase for more than 10 epochs. The model weights with the best performance on the validation set were selected as the final model. This model was then evaluated on the four independent test sets, which were not considered for training, hyper-parameter tuning or model selection.

Table 4 presents the validation and test set results of the overall MOS prediction compared to the single-ended models P.563 [5], ANIQUE+ [36], WAWEnets [12], and the double-ended models POLQA [4], DIAL [37], and VISQOL (v3.1.0) [38][3]. NISQA outperforms the other single-ended speech quality models on most of the datasets, except for the ITU-T Suppl. 23 datasets, where ANIQUE+ achieved the best results. However, it should be noted that these datasets were available for the development of ANIQUE+. POLQA outperforms NISQA on most of the POLQA-Pool datasets (103–603) that contain typical ITU-T P.800 double sentences with a silent pause in between. In contrast, NISQA achieves better results than POLQA on the NISQA test datasets that contain conversational speech.

Table 5 shows the results of the speech quality dimensions prediction for the datasets for which subjective speech quality dimension ratings are available. NISQA outperforms the double-ended model DIAL on most of the datasets and achieves overall good results with RMSEs of 0.16–0.56.

6. Conclusions

We presented the speech quality model NISQA, which, besides overall MOS, also predicts the four speech quality dimensions Noisiness, Coloration, Discontinuity, and Loudness. With this degradation decomposition approach, more insights into the cause of an underlying quality impairment are provided. The model is based on a CNN with following Self-Attention network for time-dependency modelling and an Attention-Pooling block for final time pooling. The model is trained and evaluated on a large set of 81 datasets from different sources and showed to give reliable results on unknown data and real, live phone calls. Furthermore, we open-source the code, the model weights, and speech quality datasets. The presented model is focused on distortions that occur in modern speech communication networks. However, the model weights can also be used to fine-tune the model for related tasks, such as the prediction of enhanced or synthesised speech as shown in [39].

[3] P.563 and ANIQUE+ only allow to predict narrowband signals. Therefore, all signals have been downsampled to 8 kHz sample rate for the prediction with these models. VISQOL was applied in speech mode (as it gave better results than the audio mode) that only considers frequencies up to 8 kHz (wideband). Dataset NISQA_TEST_LIVETALK contains no reference signals and can therefore only be compared to single-ended models.

7. References

[1] ITU-T Rec. P.800, *Methods for subjective determination of transmission quality*, 1996.

[2] ITU-T Rec. P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*, 2018.

[3] B. Naderi and R. Cutler, "An open source implementation of ITU-T Recommendation P.808 with validation," in *Proc. Interspeech 2020*, 2020.

[4] ITU-T Rec. P.863, *Perceptual objective listening quality assessment*, 2018.

[5] ITU-T Rec. P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, 2004.

[6] A. Hines, E. Gillen, and N. Harte, "Measuring and monitoring speech quality for voice over IP with POLQA, viSQOL and P.563," in *Proc. Interspeech 2015*, 2015.

[7] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech.* Springer, 2012.

[8] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of loudness level on the overall quality of transmitted speech," in *Proc. 123rd Audio Engineering Society Convention*, 2007.

[9] M. Soni and H. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *Proc. 2016 24th European Signal Processing Conference (EUSIPCO)*, 2016.

[10] J. Ooster and B. T. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *Proc. ICASSP 2019*, 2019.

[11] 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 *Proc. Interspeech 2018*, 2018.

[12] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *Proc. ICASSP 2020*, 2020.

[13] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. ICASSP 2020*, 2020.

[14] ——, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," in *Proc. Interspeech 2020*, 2020.

[15] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," *Proc. ICASSP 2019*, 2019.

[16] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and lstm-network," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 7, pp. 1151–1163, 2019.

[17] J. Serrà, J. Pons, and S. Pascual, "Sesqa: semi-supervised learning for speech quality assessment," *ArXiv*, vol. abs/2010.00368, 2020.

[18] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. ICASSP 2019*, 2019.

[19] ——, "Quality Degradation Diagnosis for Voice Networks — Estimating the Perceived Noisiness, Coloration, and Discontinuity of Transmitted Speech," in *Interspeech 2019*, 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS 2017*, 2017.

[21] J. Berger, A. Hellenbart, R. Ullmann, B. Weiss, S. Möller, J. Gustafsson, and G. Heikkilä, "Estimation of 'quality per call' in modelled telephone conversations," in *Proc. ICASSP 2008*, 2008.

[22] ITU-T P Suppl. 23, *Perceptual objective listening quality assessment*, 2008.

[23] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. W. Lewis, A. Butcher, and J. Hajek, "Building an audio-visual corpus of Australian English: Large corpus collection with an economical portable and replicable black box," in *Proc. Interspeech 2011*, 2011.

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

[25] P. Kabal, "TSP speech database," McGill University, Quebec, Canada, Tech. Rep. Database Version 1.0, 2002.

[26] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[27] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the english accents in the british isles," in *Proc. 12th Language Resources and Evaluation Conference (LREC)*, 2020.

[28] J. F. Gemmeke, D. P. W. 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 *Proc. ICASSP 2017*, 2017.

[29] Freesound, https://freesound.org/, 2020.

[30] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *Journal of the Acoustical Society of America*, vol. 133, pp. 3591–3591, 2013.

[31] ITU-T Rec. P.501, *Test signals for use in telephony and other speech-based application*, 2020.

[32] G. Morrison, P. Rose, and C. Zhang, "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice," *Australian Journal of Forensic Sciences*, vol. 44, pp. 155 – 167, 2012.

[33] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow. (2015) Forensic database of voice recordings of 500+ australian english speakers. [Online]. Available: http://databases.forensic-voice-comparison.net/

[34] L. Fernández Gallardo and B. Weiss, "The Nautilus Speaker Characterization Corpus: Speech recordings and labels of speaker characteristics and voice descriptions," in *Proc. of International Conference on Language Resources and Evaluation (LREC)*, 2018.

[35] G. Mittag, S. Zadtootaghaj, T. Michael, B. Naderi, and S. Möller, "Bias-aware loss for training image and speech quality prediction models from multiple datasets," in *Accepted at QoMEX 2021*, 2021.

[36] D. Kim and A. Tarraf, "ANIQUE+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, Spring 2007.

[37] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality.* Springer, 2011.

[38] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.

[39] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech 2020*, 2020.