INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Quality Degradation Diagnosis for Voice Networks – Estimating the Perceived Noisiness, Coloration, and Discontinuity of Transmitted Speech

Gabriel Mittag[1], *Sebastian Möller*[1,2]

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

Abstract

We present a single-ended quality diagnosis model for super-wideband speech communication networks, which predicts the perceived *Noisiness*, *Coloration*, and *Discontinuity* of transmitted speech. The model is an extension to the single-ended speech quality prediction model NISQA and can additionally indicate the cause of quality degradation. Service providers can use the model independently of the communication system's technology since it is based on universal perceptual quality dimensions. The prediction model consists of a convolutional neural network that firstly calculates per-frame features of a speech signal and subsequently aggregates the features over time with a recurrent neural network, to estimate the speech quality dimensions. The proposed diagnosis model achieves promising results with an average RMSE* of 0.24.
Index Terms: speech quality, deep learning, single-ended

1. Introduction

The perceived speech quality is one of the most important performance indicator for the evaluation of speech communication networks. It is traditionally measured in auditory listening experiments, in which naïve test participants rate recorded speech signals on an *absolute category rating* (ACR) scale from 1 to 5. The average across all test participants then yields the so-called *mean opinion score* (MOS). Because this procedure is time and money consuming, instrumental methods have been developed that can automatically predict the MOS value. The current recommendation by the *International Telecommunication Union* (ITU-T) for speech quality assessment is P.OLQA [1], which compares the clean reference signal with the degraded output signal to estimate the perceived speech quality. However, the clean reference signal is not always available, especially when considering online monitoring scenarios. Recently, we presented a single-ended speech quality assessment model NISQA[1] (Non-Intrusive Speech Quality Assessment) [2] that predicts the perceived speech quality without the need for a reference. Also, in contrast to other state-of-the-art single-ended speech quality models (e.g. P.563 [3] or ANIQUE+ [4]), NISQA is not restricted to narrowband (up to 4 kHz) signals but is able to predict the quality of wideband (up to 8 kHz) and super-wideband (up to 16 kHz) channels. This is important to note since narrowband calls are becoming less relevant and wideband (e.g. in over-the-top services, such as Whatsapp, Line, and Skype; or in many 3G / 4G networks with the AMR-WB codec) and super-wideband (e.g. through VoLTE/VoWiFi and the EVS codec) are becoming more popular quickly.

However, instrumental speech quality models are only able to estimate the overall perceived quality and therefore give no indication about what caused a quality degradation. In this paper, we present an extension to the NISQA model that provides additional diagnostic information about the quality degradation. To this end, we use the approach of quality-relevant perceptual dimensions that are derived through pairwise similarity and subsequent multidimensional scaling experiments, as well as semantic differential and subsequent principal component analysis. Using these methods three perceptual dimensions were identified in [5]:

- *Noisiness*: Degradation such as background, circuit, or coding noise. It is labeled with "not noisy" and "noisy".
- *Coloration*: Degradation caused by frequency response distortions, e.g. introduced by bandwidth limitation, low bitrate codecs, or packet-loss concealment. It is labeled "uncolored" and "colored".
- *Discontinuity*: Isolated or non-stationary distortions, e.g. introduced by packet-loss or clipping. It is labeled "continuous" and "discontinuous"

The resulting quality space that is spanned by these dimensions can be mapped to the overall quality perception [6] and additionally gives the possibility to identify the cause of quality losses. It was also shown in [5] that is possible to directly annotate the quality dimensions in subjective experiments, with a method that is similar to ITU-T Rec. P.835 [7], in which noisy speech signals are rated. The advantage of using perceptual quality dimensions for degradation diagnosis is that the dimensions are independent of the technology used. Therefore, the prediction models will stay useful, even when new technologies, such as codecs or concealment algorithms, are introduced to communication systems.

In [8] the intrusive model DIAL, which estimates these quality dimensions, as well as the overall quality, was presented. Based on these activities, a new work item P.AMD (Perceptual Approaches for Multi-Dimensional Analysis) was started at the ITU-T, with the corresponding non-intrusive work item P.SAMD (Single-Ended Perceptual Approaches for Multi-Dimensional Analysis) [9]. We already presented non-intrusive prediction models for the estimation of the quality dimensions *Noisiness* and *Coloration* in [10, 11], however, these models were based on handcrafted features that – while having the advantage of being easily interpretable – proved to have limited robustness towards unknown databases. Furthermore, it was challenging to find suitable features, which rely on the degraded output signal only, to estimate the dimension *Discontinuity*. Recently, the use of deep learning for audio and speech classification and recognition tasks has become increasingly popular [12, 13, 14, 15, 16, 17, 18]. In [19], we also showed that *convolutional neural networks* (CNN) can be used to detect packet-loss concealment in speech signals, which indicates that they

[1]Available at github.com/gabrielmittag/NISQA

http://dx.doi.org/10.21437/Interspeech.2019-2636

are suitable to predict the perceived *Discontinuity* as well. Also, our single-ended speech quality model NISQA [2] is based on a CNN that firstly estimates the per-frame quality of a speech signal and subsequently uses a *long short-term memory recurrent neural network* (LSTM-network) to aggregate the per-frame quality over time, to predict the overall quality. The proposed quality degradation diagnosis model in this work is an extension of the NISQA model that uses, in principle, the same approach to predict the three quality dimensions *Noisiness*, *Coloration*, and *Discontinuity*.

2. Model

The presented model is based on a convolutional neural network that predicts the per-frame quality, based on spectrogram segments centered around the frame to be estimated. The estimated per-frame quality is then used as input to an LSTM-network, together with a CNN feature vector. A block diagram of the proposed model can be seen in Figure 1. The model input is the degraded speech signal with $f_s = 48\,\text{kHz}$. The signal is then transformed to log-Mel-spectrograms with an FFT of 1024 samples window length, a hop size of 480 samples (10 ms), and a mel filter bank with 48 bands from 0 - 16 kHz. A segment of 15 frames length (150 ms), centered around the frame to estimate the speech quality, is then extracted from the spectrogram as input to the CNN.

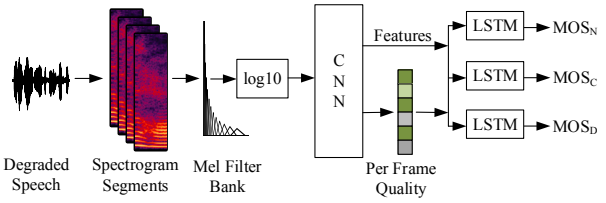


Figure 1: *Block diagram of the proposed NISQA diagnostic model*

The model follows in principle the same architecture as the model presented in [2], which predicts the overall quality. However, instead of MFCCs features together with the per-frame quality, we use features directly calculated from the CNN for the estimation of the quality dimensions. In this way, we further reduce the tendency of the network to overfit to training data, which is important, since we have fewer data available for the quality dimensions than for the training of the overall quality. To extract the features, we add another fully connected layer with output size five, just before the final fully connected layer (see the design of the CNN in Table 1).

The output of this layer is then used as input to the LSTM-network. We tried fully connected layers with different output sizes (40, 30, 20, 15, 10, and 5) but obtained the best results with a smaller output size of only 5 features. Subsequently, an LSTM-network with the six described features as input (fully connected output + per-frame quality) estimates the quality dimensions. The design of this LSTM-network is shown in Table 2. One advantage of LSTMs is that they allow signals with different lengths as input and that they are able to learn time dependencies. This property is very useful for the prediction of speech quality since short interruptions in the speech signal are often perceived as more annoying by service users than steady background noise. Simple measures, such as mean and standard deviation of per-frame based features are therefore less suitable

Table 1: *Design of the convolutional neural network*

	Layer	**Size**	**Stride**
	Conv, 16 ch	3x3	
	Batch normalization		
	ReLU		
	Maxpool	2x2	2x2
	Conv, 32 ch	3x3	
	Batch normalization		
	Relu		
2x	Maxpool	2x2	2x2
2x	Dropout 20%		
2x	Conv, 64 ch	3x3	
2x	Batch normalization		
2x	ReLU		
2x	Dropout 20%		
2x	Conv, 64 ch	3x3	
2x	Batch normalization		
2x	ReLu		
	Fully connected*, output size: 5		
	Batch normalization		
	ReLu		
	Fully connected, output size: 1		
	Regression*		

** LSTM input features*

for the prediction of speech quality. Instead of predicting all three quality dimensions with one LSTM-network, we decided to train three individual networks. Thus, we are able to balance the training data input for each dimension individually and obtain significantly better results. However, all three dimension prediction models use the same CNN network in the first step.

Table 2: *Design of the recurrent neural network*

Layer	**Units**
BiLSTM	100
Dropout 50%	
BiLSTM	125
Fully connected	
Regression	

3. Databases

In total, 15 different databases were available for training and testing of the proposed model, an overview can be seen in Table 3. Ten of the databases (1-7 & 13-15) were part of the ITU-T P.OLQA competition and contain a wide variety of different degradations, including live recordings. Three of the databases are part of the P.AMD pool (9-11) and have subjective quality dimension scores. Additionally, for training, we generated a new database (*IS19TRAIN*) with 240 different conditions, for example, clipping, different bitrate modes of the codecs G.711, G.722, AMR-NB, AMR-WB, EVS, and Opus, as well as packet-loss, white noise and other background noises, taken from [20]. The source speech files are based on the speech signals from [21], from which we took the provided two-seconds snippets of 20 talkers to generated 40 double sentences, by adding a variable pause of 0-2 seconds before the first snippet, in the middle of both snippets, and after the second snippet.

Table 3: *Overview of the databases*

Nr	Database	Lang	Con	Files Per Con	Listeners Per File
	Training				
1	ERICSSON101	sv	57	12	Expert score
2	ERICSSON103	sv	54	12	Expert score
3	OPTICOM301	cs	50	4	Expert score
4	OPTICOM302	en	44	4	Expert score
5	OPTICOM303	en	54	4	Expert score
6	PSYTECHNICS401	en	48	24	Expert score
7	PSYTECHNICS403	en	48	24	Expert score
8	IS19TRAIN	en	240	40	Expert score
9	DTAG1	de	66	12	3
10	DTAG2	de	76	12	4
11	ORANGE1	fr	56	4	18
	Test				
12	TUBDIS	de	20	2	38
13	SWISS501	de	50	4	24
14	SWISS502	de	50	4	24
15	SWISS503	de	54	4	20

For seven of the databases the quality dimensions *Noisiness*, *Coloration*, and *Discontinuity* were subjectively rated in auditory listening experiments with naïve participants. However, since the number of speech signals in these seven databases are not enough for training and testing of a deep learning based model, we annotated further databases in expert listening tests. The experts in these tests rated the conditions only, rather than each individual speech file. They listened to a selection of speech files, considered the condition description, and the per-condition overall MOS from previous subjective experiments to give a score (except for database 8, where the P.OLQA MOS scores were used). A per-file dimension MOS value was then estimated for all quality dimensions, based on the available overall per-file and per-condition MOS, with the following formula:

$$\mathrm{MOS}_{\mathrm{Dim,File}} = 0.728 - 0.692\,\frac{\mathrm{MOS}_{\mathrm{Con}}}{\mathrm{MOS}_{\mathrm{File}}} + 0.992\,\mathrm{MOS}_{\mathrm{Dim,Con}}. \quad (1)$$

The formula was derived by mapping the per-condition dimension MOS to the per-file dimension MOS of all databases for which subjective ratings are available. The estimation of a per-file MOS, rather than the assignment of each file with the per-condition dimension MOS, is especially useful for conditions with a high standard deviation between per-file ratings, such as packet-loss.

Additionally, for testing the model, we conducted auditory experiments to annotate the quality dimensions of databases 13-15 from the P.OLQA pool. Also, we generated and annotated a further database 12, which contains different packet-loss degradations of the codecs Opus and G.722, as well as the P.OLQA anchor conditions. The database contains the same German reference files as databases 13-15.

4. Training

The training of the CNN is purely based on the per-frame similarity output of the SQuadAnalyzer v.2.4.2.7 implementation of P.OLQA. P.OLQA compares the clean reference file with the degraded output file and estimates an objective MOS, additionally the implementation outputs for every frame, how similar the degraded signal is, compared to the reference signal. Thus, no subjective data is needed to train the CNN model. After aligning the per-frame similarity to the calculated spectrograms with a nearest neighbor interpolation, we train the CNN with the ADAM solver, an initial learning rate of 0.001, and mini-batch size of 4000. During the training, the CNN automatically finds a set of features that are suitable for the prediction of speech quality. We make use of this property by extracting the features just after the first fully connected layer. However, the resulting features are not focused on a specific quality dimension but rather on overall quality.

After that, we train the LSTM-networks with the estimated per-frame quality and the CNN features as inputs and the subjective and expert quality dimension scores as target values. We again use the ADAM solver, a padding value of 0, a mini-batch size of 200, and an initial learning rate of 0.001. All in- and outputs are normalized with the z-score method. Furthermore, we apply three different methods to the training data to improve the results:

- To increase the robustness of the model, we added white noises with a random level of 60 - 100 dB to the training speech signals, prior to calculating the CNN features and estimating the per-frame quality, with the assumption that noise in this magnitude range does not affect perceived quality.
- Because many conditions in the databases affect only one of the three dimensions, our training data is highly unbalanced with a large number of high-quality ratings. Therefore, to avoid biased training, we balance the training data by omitting a certain percentage of speech signals that have a high-quality rating (e.g. $\mathrm{MOS} > 4$).
- Speech quality ratings are often subject to a database specific bias. A constant or linear bias in a database can, for example, be caused by the overall quality that is contained in a database and thus influence the participants to score more pessimistically or more optimistically. Also, different talkers, file content and the test equipment can have an influence on the participants quality perception [22]. Because of this, we normalize each database by setting the MOS value of the clean reference condition to $\mathrm{MOS} = 4.75$ with the following formula (thus, effectively assuming a linear bias):

$$\mathrm{MOS}_{\mathrm{Norm}} = \frac{\mathrm{MOS}-1}{\mathrm{MOS}_{\mathrm{Clean}}-1}\,3.75 + 1. \quad (2)$$

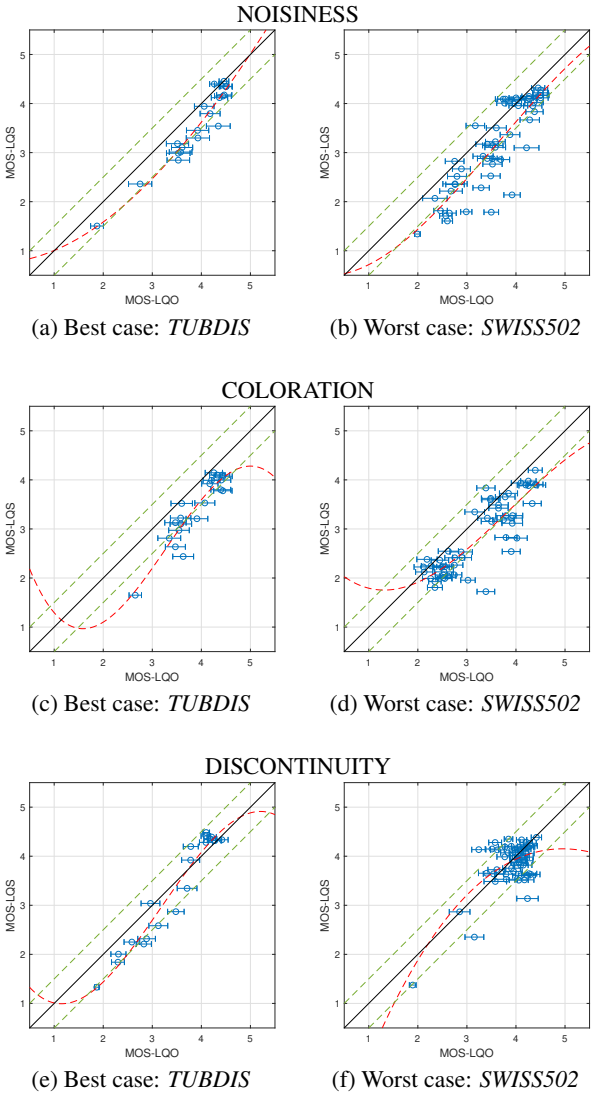
5. Results

The results are evaluated in terms of the *epsilon-insensitive RMSE* RMSE* after the application of a third order polynomial monotonous mapping, according to ITU-T Rec. P.1401 [22]. The calculation of the RMSE* is similar to the traditional *root mean square error* (RMSE) but takes into account the confidence interval of the individual MOS scores (see P.1401 eq. (7.29)). The mapping compensates for offsets, different biases, and other shifts between scores from the individual experiments, without changing the rank order. Additionally, we include the Pearson correlation coefficient *r* to the evaluation analysis, although it is not used as performance criteria within the ITU-T. The performance objective in the ITU-T P.AMD requirement specification [9] is defined as an average RMSE* smaller than 0.35 and a worst case RMSE* of 0.5 for unknown test databases. Additionally, we compare the performance with the intrusive model DIAL, which – in contrast to the proposed model – also uses the clean reference signal for the quality prediction.

The results of the test databases are presented in Table 4. It can be seen that, on average, the proposed single-ended model NISQA outperforms the intrusive model DIAL for all three

Table 4: *Results on the test databases (condition based), compared to the intrusive model DIAL*

	NOISINESS				COLORATION				DISCONTINUITY			
	NISQA		DIAL		NISQA		DIAL		NISQA		DIAL	
	r	RMSE*	*r*	RMSE*	*r*	RMSE*	*r*	RMSE*	*r*	RMSE*	*r*	RMSE*
TUBDIS	0.96	0.12	0.86	0.25	0.93	0.13	0.87	0.14	0.97	0.14	0.87	0.34
SWISS501	0.89	0.34	0.81	0.47	0.92	0.14	0.75	0.29	0.77	0.24	0.73	0.36
SWISS502	0.86	0.35	0.69	0.50	0.85	0.27	0.81	0.33	0.68	0.32	0.55	0.36
SWISS503	0.87	0.22	0.84	0.25	0.87	0.22	0.89	0.20	0.73	0.32	0.73	0.32
Average	0.90	0.26	0.80	0.37	0.89	0.19	0.83	0.24	0.79	0.26	0.72	0.35



(a) Best case: *TUBDIS* (b) Worst case: *SWISS502*

(c) Best case: *TUBDIS* (d) Worst case: *SWISS502*

(e) Best case: *TUBDIS* (f) Worst case: *SWISS502*

Figure 2: *Correlation diagram of the best and worst case results (condition based). The error bars indicate the 95% confidence interval, the red dashed line represents the 3rd order mapping*

quality dimensions. The highest error of all dimensions and databases is an RMSE* of 0.35 and thus well below the specified performance objective of 0.5. The highest error on average is obtained for the quality dimensions *Noisiness* and *Discontinuity* with an RMSE* of 0.26, which is also well below the performance objective of 0.35. The quality dimension with the lowest error is *Coloration* with an RMSE* of 0.19, it also is the dimension for which the DIAL model obtains the best results and thus, seems to be the least challenging to predict. In fact, the largest influence on the *Coloration* is the audio bandwidth limitation, which can be estimated rather straightforwardly. However, for example, non-linear distortions of codecs and robotic voices, caused by packet-loss concealment, also have an influence on the *Coloration* and are more difficult to detect in the signal. On the other hand, the resulting correlations of the *Discontinuity* dimension are relatively low, while the RMSE* values are still good. This effect can be observed for the NISQA, as well as the DIAL results, and is partly caused by the unbalanced *Discontinuity* MOS values in the databases. Figure 2 shows the best and worst case results in correlation diagrams between the estimated MOS and the subjective MOS values. The red dashed line represents the third order polynomial monotonous mapping. In the figure, it can be seen that most of the subjective MOS values for the *Discontinuity* range between 3.5 and 4.5, which results in low correlation scores. The TUBDIS database, which concentrates on packet-loss, has a better distribution of *Discontinuity* MOS values and thus yields better Pearson correlation values.

Generally, it can also be observed that the NISQA model tends to overrate the quality. This could be a result of the normalization step that was applied to the training data since clean reference signals often obtain lower MOS scores in subjective experiments than the set value of 4.75.

6. Conclusions

In this paper, we presented a single-ended quality degradation diagnosis model, which can be used for degradation diagnosis of super-wideband speech communication systems. It uses a deep learning approach with a CNN-LSTM-network to predict the three quality dimensions *Noisiness*, *Coloration*, and *Discontinuity*. These quality dimensions are directly linked to technical root-causes and can thus give an indication of what caused a quality degradation in the communication system. The proposed model is well below the performance objectives, specified by the ITU-T, with an average RMSE* across all dimensions of 0.24 and a worst-case RMSE* of 0.35. Furthermore, it outperforms the intrusive diagnostic model DIAL. In future work, we plan to use spontaneous speech as training and test data, which is more realistic for single-ended scenarios. Also, we plan to carry out more subjective experiments to improve the results and rely less on expert scores.

7. Acknowledgements

The work on this paper was largely supported by the BMBF, Grant 01IS17052 and by the Deutsche Forschungsgemeinschaft, Grant MO 1038/22-1.

8. References

[1] ITU-T Rec. P863, "Perceptual objective listening quality assessment."

[2] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," Accepted for publication in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[3] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Geneva.

[4] 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.

[5] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*. Berlin, Heidelberg: Springer, 2012.

[6] F. Köster, G. Mittag, and S. Möller, "Modeling the overall quality of experience on the basis of underlying quality dimensions," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[7] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm."

[8] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Berlin, Heidelberg: Springer, 2011.

[9] ITU-T SG 12 (Study Period 2017) Temporary Document 137-GEN, "Technical Requirement Specification P.AMD and P.SAMD."

[10] F. Köster, G. Mittag, T. Polzehl, and S. Möller, "Non-intrusive Estimation of Noisiness as a Perceptual Quality Dimension of Transmitted Speech," in *5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, 2016, pp. 74–78.

[11] G. Mittag, F. Köster, and S. Möller, "Non-intrusive estimation of the perceptual dimension coloration," in *Fortschritte der Akustik, DAGA 2016: Plenarvortr. u. Fachbeitr. d. 42. Dtsch. Jahrestg. f. Akust.*, 2016, pp. 976–979.

[12] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6979–6983.

[13] J. Pons and X. Serra, "Designing efficient architectures for modeling temporal features with convolutional neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2472–2476.

[14] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 559–563.

[15] L. Tóth, "Phone recognition with hierarchical convolutional deep maxout networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–13, 2015.

[16] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings INTERSPEECH*, 2017, pp. 3512–3516.

[17] J. Ooster, R. Huber, and B. T. Meyer, "Prediction of perceived speech quality using deep machine listening," in *Proc. Interspeech 2018*, 2018, pp. 976–980.

[18] S. wei 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, pp. 1873–1877.

[19] G. Mittag and S. Möller, "Non-intrusive estimation of packet loss rates in speech communication systems using convolutional neural networks," in *Proceedings of IEEE International Symposium on Multimedia (ISM)*, 2018, pp. 105–109.

[20] ETSI EG 202 396-1, "Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database."

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

[22] ITU-T Rec. P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models."