NON-INTRUSIVE SPEECH QUALITY ASSESSMENT FOR SUPER-WIDEBAND SPEECH COMMUNICATION NETWORKS

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

[1] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2] Language Technology, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany

ABSTRACT

The quality of speech communication networks has recently improved significantly by extending the available audio bandwidth from narrowband, firstly to wideband, and then to super-wideband. This bandwidth extension marks the end of the typically muffled sound we know from plain old telephone services. Another reason for increased speech quality is the fully digitally packet-based transmission. However, so far, no speech quality prediction model is able to estimate super-wideband quality without a clean reference signal. In this paper, we present a non-intrusive speech quality assessment model NISQA, which – in contrast to current state-of-the-art models – can predict the quality of super-wideband speech transmission. Furthermore, it is able to accurately predict the quality impact of packet loss concealment of modern codecs, such as Opus and EVS. The model uses a novel approach, where a CNN firstly estimates the per-frame quality, and subsequently, an RNN aggregates the per-frame values over time, to estimate the overall speech quality. Averaged over a comprehensive test set, the model achieves an RMSE*3rd of 0.29 with subjective MOS.

Index Terms— speech quality, non-intrusive, single-ended, packet loss, quality of service

1. INTRODUCTION

The perceived speech quality of voice communication services has rapidly increased in the last decade. One of the reasons for the improved quality is the extension of the transmission bandwidth from *narrowband* (NB), with a bandwidth from 300 - 3400 Hz, to wideband (WB) with 100 - 7000 Hz. These days, an increasing amount of telephone calls are carried out in WB, for example in mobile 3G or in fixed-line *Voice over IP* (VoIP) networks. Recently, the quality was further improved with the introduction of super-wideband (SWB) transmission to speech communication networks, with a bandwidth of 50 - 14000 Hz. In mobile voice networks, SWB is enabled with the state-of-the-art codec EVS and *Voice over LTE* (VoLTE) or *Voice over WLAN* (VoWi-Fi) technology. Also, many over-the-top VoIP services (e.g. Whatsapp, Skype, Line, etc.) support WB/SWB transmission with codecs such as Opus.

The quality of speech transmission services is traditionally assessed in *listening-only test*, in which naïve test participants judge the quality of speech samples on a 5-point *absolute category rating* (ACR) scale. The average across all test participants then gives the *mean opinion score* (MOS). Since subjective methods require a significant effort to conduct, instrumental models have been established. Signal-based models can be divided into two groups: *Intrusive* models require the degraded output signal of the transmission system *and* the clean original input signal. The quality is then estimated with the help of distance measurements between both signals. *Non-intrusive* or single-ended models rely only on the degraded output signal of the transmission system. The long-term standard for NB and WB speech quality assessment by the *International Telecommunication Union* (ITU-T) has been PESQ [1] and WB-PESQ [2]. They are now replaced by P.OLQA [3], the current recommendation by the ITU-T, which also considers SWB transmission. P.OLQA proved to deliver reliable predictions if the reference signal is available.

However, the currently recommended non-intrusive model by the ITU-T P.563 [4] only considers NB transmission. To the best of our knowledge, so far, no non-intrusive speech quality models for WB nor SWB transmission have been proposed. That being said, many NB models have been presented in the literature. Apart from P.563, the ANIQUE+ model [5] also showed to provide accurate prediction results. More recently, in [6], a model based on the outputs of an automatic speech recognizer was presented, and in [7] a model based on a BiLSTM network was shown, which is focused towards speech enhancement. The latter paper also includes a comprehensive list of other proposed models. The current state of the art models P.563 and ANIQUE+ are known to give poor quality estimates in case of concealed packet loss. However, in modern packet-based communication networks, where the speech transmission is fully digitally, lost packets are one of the main quality impairments. In this paper, we try to overcome these problems by presenting a model that predicts the quality of transmitted speech in an SWB context.

The work on this paper was largely supported by the BMBF, Grant 01IS17052.

2. MODEL DESCRIPTION

The proposed SWB speech quality estimator NISQA[1] is based on a *convolutional neural network* (CNN) that estimates the speech quality for each frame of the input signal. The estimated per-frame quality values are then aggregated over time by using a *recurrent neural network* (RNN). The input to our model are signals with a sample rate of 48 kHz, which are then transformed to log-mel-spectrograms (see Figure 2). To do this, firstly, we calculate spectrograms with an FFT window length of 1024 samples. We use a hop size of 480 samples to obtain a time resolution of 10 ms. Then, a segment of the spectrogram with length 15 frames (i.e. 150 ms) is extracted, centered around the frame to estimate the speech quality. After this step, a mel filter bank with a frequency range from 0 - 16 kHz and 48 bands is applied. The log energy is then used as input for the CNN. CNNs are

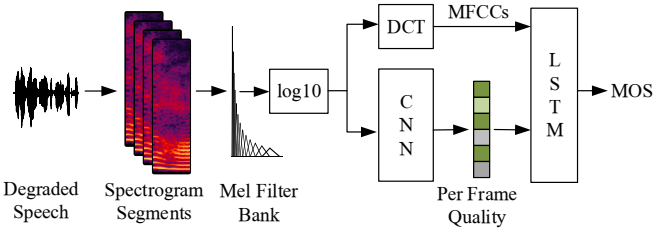


Fig. 1. Block diagram of the proposed NISQA model.

most commonly used in the field of image classification and have the ability to learn a suitable set of features – for a given regression or classification task – automatically. This feature set learning is done in the so-called convolutional layers that give the neural network its name. Recently, they have increasingly been used for recognition or detection tasks in the context of music and speech [8, 9, 10, 11, 12]. In [13] we further showed that CNNs can successfully be used to detect lost packets from speech spectrograms. The design of the proposed CNN is shown in Table 1. We use three maxpooling layer to downsample the feature map in time and frequency. The downsampling procedure helps to fasten computation time and also avoids overfitting of the model by reducing the number of parameters in the network. To further reduce overfitting, dropout layers [14] are applied. To speed up the training time we use batch normalization [15] and *rectified linear unit* (ReLU) layers. The output of the CNN is then the per-frame quality. It should be noted that it is not possible to model the overall quality by using simple metrics such as the average and variance of the estimated per-frame quality. For example, short interruptions have been proven to sound more annoying than steady background noise [16]. Also, low quality, during a silent segment of the original speech signal,

[1]The proposed model is available to download for other researchers at: https://github.com/gabrielmittag/NISQA

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		
	Maxpool	2x2	2x2
	Dropout 20%		
	Conv, 64 ch	3x3	
	Batch normalization		
2x	ReLU		
	Dropout 20%		
	Conv, 64 ch	3x3	
	Batch normalization		
	Regression		
	Fully connected		
	Softmax		

degrades the overall speech quality less than during active speech segments, where the impairment may even disturb the intelligibility of the transmitted speech. For these reasons, we chose to model the time dependency of the perceived speech quality with an RNN. Another advantage of RNNs is that they allow time sequences with different lengths as input. *Long short-term memory* (LSTM) networks are a type of RNN that is able to remember their inputs over a longer period of time and thus are able to model long-term dependency problems. These kinds of networks are often used to forecast time-series. A more specific variation is the BiLSTM layer that learns bidirectional dependencies between time steps and which we used in our model. The design of the RNN that is used to estimate the overall speech quality is depicted in Table 2. Besides the per-frame quality, we also use *mel-frequency ceptral coefficients* (MFCCs) as inputs, in order to provide the RNN with some context of the speech signal. We use MFCCs over the mel filter bank output because they compress information about the vocal tract into a small number of coefficients. By using a more compressed feature set, we hope to avoid overfitting of the LSTM network. To calculate the MFCCs we apply a *discrete cosine transform* (DCT) to the log-mel-spectrogram and extract 13 coefficients. The advantage of the CNN-LSTM approach is twofold: Firstly, the per-frame quality gives some insight into the cause of a quality degradation, for example, frame loss due to transmission errors could be observed by a sudden decrease of speech quality. Secondly, this approach helps to regularize the training of the RNN. As on a per frame basis, we have a large amount of training data, but on a file basis only limited data, it is important to minimize the input feature size of the RNN.

3. DATABASES

Overall, 29 different databases with typical P.800 [17] double sentences with a duration of 6 - 12 s were available. 20 of the databases are taken from the P.OLQA pool and were used during the ITU-T P.OLQA competition. These databases contain a large variety of speech distortions, such as different codecs, noises, live recordings, and transmission errors. All SWB test set databases from the P.OLQA pool were chosen for our test set and all SWB training sets were included in our training set. Additionally, we included the WB databases, for which SWB reference files were available and calculated an SWB MOS with P.OLQA for training only; the same was done for the additional databases *WB_DTAG1* and *WB_DTAG2*, which are described in [16]. Three more SWB sets were included: *SWB_TUBDIS* contains the same anchor conditions as the databases of the P.OLQA pool and additionally packet loss conditions of the codecs G.722 and OPUS, *SWB_TUBLIK* contains different codecs, band-passes, and MNRU noise and is described in [18], *SWB_VUPL* contains different packet loss conditions of the EVS codec at the highest bitrate and uses the German ITU-T P.501 [19] Annex C sentences as source files.

Many of the databases are using the same reference signals. However, for non-intrusive models, it is important to validate that the model gives accurate predictions for a wide variety of talkers. Also, more recent SWB codecs, such as Opus and EVS are not included in the P.OLQA pool. Because of this, we generated four databases with reference files taken from the TSP database [20] and overall 16 different talkers, where each of the four databases is generated from 120 sentences by four talkers. Then we processed the reference files with the codecs G.711, G.722, AMR-NB, AMR-WB, OPUS, and EVS in different bitrate modes and random and bursty packet loss from 1-12 %. The packet loss patterns were applied with the ITU-T STL Toolbox [21], where bursty error patterns are based on the Bellcore model [22] and random error patterns on the Gilbert model. The Opus decoder was modified in order to apply error patterns from the STL Toolbox. These four databases are not subjectively rated and only the objective MOS, calculated by P.OLQA is available. An overview of the language, number of conditions, files, listeners, and talkers of the databases can be seen in Table 3.

Table 2. Design of the recurrent neural network.

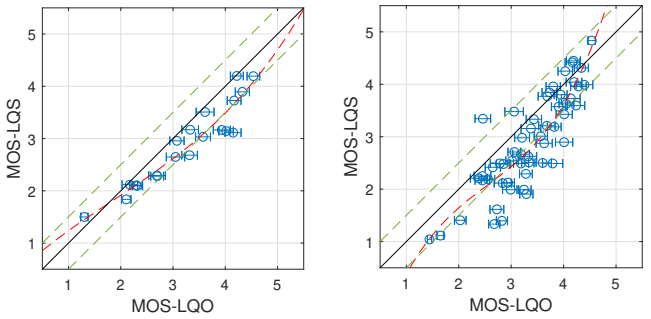
Layer	Units
BiLSTM	100
Leaky ReLU	
Dropout 50%	
BiLSTM	125
Fully connected	
Regression	

4. MODEL TRAINING AND RESULTS

To train the model, we first calculated the per-frame similarity between the degraded and the original signal with POLQA v2 in the SQuadAnalyzer v.2.4.2.7 implementation. Then we aligned the per-frame similarity with the spectrogram segments, using a nearest neighbor interpolation. The aligned similarity is then used as the response variable for the CNN training. We use the ADAM solver, a mini-batch size of 4000 and an initial learning rate of 0.001. After the CNN training, we use the prediction of the per-frame quality/similarity of the training set, together with the MFCCs as input for the RNN training. To this end, we again use the ADAM solver, a mini-batch size of 200, padding value of 0, and an initial learning rate of 0.001. We shuffled the mini-batch at every epoch. For the evaluation, however, we estimated the responses individually for every file, without padding. All in- and outputs are normalized with the z-score method. As response variable, we applied the subjective per-file MOS for the SWB databases, and the objective P.OLQA SWB MOS for the WB databases and the *SWB_TSP_PL* databases, for which no subjective ratings are available. The results are evaluated according to ITU-T Rec. P.1401 [23] in terms of the *epsilon-insensitive RMSE* (RMSE*) after a 3rd order polynomial monotonic mapping (RMSE*3rd). The RMSE* is similar to the traditional *root mean square error* (RMSE) but considers 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* and the traditional RMSE to the evaluation analysis. To compare our results to other non-intrusive speech quality models we used the current state-of-the-art models ANIQUE+ and P.563. Both of these models are only valid for narrowband signals with a sample rate of 8 kHz. However, since there are no WB or SWB models available yet, we downsampled the speech signals in order to have a baseline for comparison. We used the objective MOS together with the calculated features to retrain both models on the training set with a linear regression. It should be noted that the retrained versions improve the results of ANIQUE+ and P.563 on all databases. In addition, we compare the results to the intrusive model P.OLQA, which can be seen as the topline with the lowest RMSE*3rd possible. The results are presented in Table 3. The average is calculated only over databases with subjective scores. On the training set, the model achieves excellent results. This was expected since we use neural networks that tend to overfit to training data. The average RMSE*3rd on the test data is 0.29, which is approximately 0.1 higher than on the training data. However, the proposed model NISQA still outperforms the retrained baseline models ANIQUE+ and P.563 by an RMSE*3rd of 0.11 on average. The RMSE*3rd of the NISQA model is on average only 0.1 worse than the

Table 3. Results of the proposed NISQA model compared to P.OLQA, and retrained Anique+, and P.563.

Databases	Lang	Con	Files Per Con	Listeners Per File	Source Talker	P.OLQA			ANIQUE+ RT			P563 RT			NISQA		
						r	RMSE	RMSE*3rd	*r*	RMSE	RMSE*3rd	*r*	RMSE	RMSE*3rd	*r*	RMSE	RMSE*3rd
Training																	
SWB_101_ERICSSON	sv	57	12	10	a	0.83	0.56	0.24	0.52	0.62	0.52	0.66	0.54	0.44	0.92	0.37	0.19
SWB_201_FT_DT	fr	48	4	24	b	0.93	0.44	0.27	0.68	0.78	0.59	0.76	0.70	0.50	0.92	0.57	0.25
SWB_202_FT_DT	fr	49	4	24	b	0.84	0.60	0.24	0.64	0.83	0.62	0.77	0.62	0.50	0.93	0.36	0.23
SWB_301_OPTICOM	cs	50	4	24	c	0.91	0.37	0.24	0.63	0.73	0.61	0.70	0.65	0.54	0.88	0.54	0.32
SWB_302_OPTICOM	en	44	4	24	d	0.93	0.47	0.15	0.47	0.79	0.59	0.64	0.64	0.49	0.89	0.41	0.24
SWB_401_PSYTECHNICS	en	48	24	8	e	0.96	0.27	0.13	0.88	0.56	0.28	0.87	0.51	0.34	0.97	0.30	0.15
SWB_501_SWISSQUAL	de	50	4	24	f	0.92	0.32	0.24	0.38	0.75	0.66	0.66	0.62	0.50	0.94	0.31	0.17
SWB_502_SWISSQUAL	de	50	4	24	f	0.90	0.42	0.25	0.69	0.67	0.50	0.69	0.61	0.50	0.91	0.43	0.23
SWB_601_TNO	nl	50	4	24	g	0.95	0.39	0.22	0.55	0.80	0.67	0.70	0.66	0.56	0.95	0.41	0.20
SWB_602_TNO	nl	50	4	24	g	0.96	0.32	0.17	0.53	0.77	0.68	0.64	0.71	0.59	0.96	0.25	0.15
SWB_GIPS_EXP4	en	36	4	25	h	0.93	0.54	0.07	0.77	0.44	0.20	0.78	0.34	0.17	0.94	0.37	0.07
WB_DTAG1	de	66	12	3	i	0.96	0.65	0.08	0.41	0.78	0.61	0.66	0.63	0.41	0.95	0.76	0.14
WB_DTAG2	de	76	12	4	i	0.90	0.39	0.16	0.47	0.63	0.48	0.78	0.53	0.29	0.83	0.59	0.21
WB_204_FT_DT	fr	53	4	24	b	0.85	0.48	0.22	0.66	0.70	0.53	0.86	0.53	0.30	0.91	0.41	0.26
WB_402_PSYTECHNICS	en	48	24	8	e	0.98	0.20	0.16	0.86	0.74	0.39	0.84	0.64	0.39	0.96	0.47	0.18
WB_102_ERICSSON	sv	54	12	13	a	0.87	0.33	0.17	0.46	0.58	0.48	0.60	0.58	0.42	0.83	0.39	0.26
SWB_TSP_PL_A	en	245	4	-	j	-	-	-	0.50	0.76	-	0.63	0.71	-	0.95	0.40	-
SWB_TSP_PL_B	en	245	4	-	k	-	-	-	0.50	0.74	-	0.64	0.68	-	0.95	0.35	-
Test																	
SWB_103_ERICSSON	sv	54	12	8	a	0.90	0.42	0.24	0.64	0.59	0.45	0.77	0.47	0.35	0.83	0.56	0.29
SWB_203_FT_DT	fr	54	4	24	b	0.86	0.49	0.29	0.49	0.81	0.72	0.70	0.70	0.56	0.85	0.67	0.37
SWB_303_OPTICOM	en	54	4	24	d	0.92	0.45	0.16	0.70	0.84	0.48	0.87	0.63	0.30	0.85	0.44	0.33
SWB_403_PSYTECHNICS	en	48	24	8	e	0.98	0.23	0.16	0.79	0.61	0.41	0.86	0.49	0.34	0.89	0.61	0.29
SWB_503_SWISSQUAL	de	54	4	24	f	0.93	0.34	0.18	0.63	0.73	0.58	0.76	0.64	0.45	0.83	0.49	0.37
SWB_603_TNO	nl	48	4	24	g	0.97	0.26	0.16	0.67	0.76	0.59	0.82	0.63	0.43	0.88	0.49	0.33
SWB_TUBDIS	de	20	2	41	f	0.94	0.56	0.15	0.80	0.49	0.37	0.80	0.48	0.37	0.95	0.47	0.16
SWB_TUBLIK	de	8	12	20	l*	0.99	0.28	0.16	0.85	0.89	0.74	0.92	0.75	0.51	0.98	0.25	0.18
SWB_TUBVUPL	de	15	4	36	m*	0.70	0.63	0.26	0.81	0.64	0.37	0.85	0.47	0.32	0.88	0.60	0.29
SWB_TSP_PL_C	en	245	4	-	n*	-	-	-	0.45	0.75	-	0.58	0.70	-	0.92	0.46	-
SWB_TSP_PL_D	en	245	4	-	o*	-	-	-	0.54	0.73	-	0.64	0.68	-	0.92	0.43	-
Average (subj. data only)																	
Training						0.91	0.42	0.19	0.60	0.70	0.53	0.72	0.59	0.43	0.92	0.43	0.20
Test						0.91	0.41	0.19	0.71	0.71	0.52	0.82	0.59	0.40	0.89	0.50	0.29



(a) Best case: *SWB_TUBDIS* (b) Worst case: *SWB_203_FTDT*

Fig. 2. Correlation diagram of the best and worst case results. The error bars indicate the 95% confidence interval, the red dashed line represents the 3rd order mapping.

one of the reference based model P.OLQA. The worst case RMSE*3rd is 0.37 and the best RMSE*3rd 0.16 is obtained for set *SWB_TUBDIS*, which contains different packet loss conditions for the codecs G.722 and Opus. The results for the best and worst case are shown in Figure 2.

Current non-intrusive speech quality models are known for underestimating the quality impairment impact of packet loss concealment algorithms. It can be seen that the prediction accuracy of the two baseline models for the databases *SWB_TSP_PL_C* and *SWB_TSP_PL_D*, which focus on packet-loss, is poor. The proposed model, however, has a high correlation of $r = 0.92$ to the results obtained by P.OLQA for these two databases. Furthermore, the results for the databases with unknown talkers (marked with *) are not worse than for known talkers, which indicates that the model delivers good predictions, independent of the talker or sentence.

5. CONCLUSION

We presented a new non-intrusive speech quality assessment model NISQA for SWB transmission. The model is based on a novel approach, where a CNN is used to estimate the per-frame quality, and an RNN aggregates the per-frame values over time to predict the overall speech quality. We showed that the proposed model is able to give good prediction results over the same test set that was used for the P.OLQA validation, with an average RMSE*3rd of 0.29 and a worst-case RMSE*3rd of 0.37. Furthermore, we showed that the model can be used across different talkers and sentences. In contrast to the current state-of-the-art models, NISQA is able to predict the speech quality of packet loss concealment conditions of modern speech codecs. Also, the model is made available on a GitHub for research purposes. In the future, we will extend our approach to a diagnostic model that estimates the perceptual quality dimensions noisiness, coloration, discontinuity, and loudness. This work is then planned to be contributed to the work item P.SAMD of the ITU-T SG12.

6. REFERENCES

[1] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," .

[2] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," .

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

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

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

[6] Jasper Ooster, Rainer Huber, and Bernd T. Meyer, "Prediction of perceived speech quality using deep machine listening," in *Proc. Interspeech 2018*, 2018, pp. 976–980.

[7] Szu wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech 2018*, 2018, pp. 1873–1877.

[8] 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.

[9] 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.

[10] 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.

[11] László 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.

[12] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings INTERSPEECH*, 2017, pp. 3512–3516.

[13] G. Mittag and S. Möller, "Non-intrusive estimation of packet loss rates in speech communication systems using convolutional neural networks," Accepted for publication in *Proc. 2018 IEEE International Symposium on Multimedia (ISM)*, 2018.

[14] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

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

[17] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," .

[18] L. F. Gallardo, G. Mittag, S. Möller, and J. Beerends, "Variable voice likability affecting subjective speech quality assessments," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–6.

[19] ITU-T Rec. P.501, "Test signals for use in telephonometry," .

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

[21] ITU-T Rec. G.191, "ITU-T Software Tool Library 2009 User's manual," .

[22] V. K. Varma, "Testing speech coders for usage in wireless communications systems," in *Proceedings., IEEE Workshop on Speech Coding for Telecommunications,*, 1993, pp. 93–94.

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