DNSMOS P.835: A NON-INTRUSIVE PERCEPTUAL OBJECTIVE SPEECH QUALITY METRIC TO EVALUATE NOISE SUPPRESSORS

Chandan K A Reddy, Vishak Gopal, Ross Cutler

Microsoft Corporation, Redmond, WA
chandan.ka@outlook.com, vishak.gopal@microsoft.com, ross.cutler@microsoft.com

ABSTRACT

Human subjective evaluation is the "gold standard" to evaluate speech quality optimized for human perception. Perceptual objective metrics serve as a proxy for subjective scores. We have recently developed a non-intrusive speech quality metric called Deep Noise Suppression Mean Opinion Score (DNSMOS) using the scores from ITU-T Rec. P.808 [1] subjective evaluation. The P.808 scores reflect the overall quality of the audio clip. ITU-T Rec. P.835 [2] subjective evaluation framework gives the standalone quality scores of speech and background noise in addition to the overall quality. In this work, we train an objective metric based on P.835 human ratings that output 3 scores: i) speech quality (SIG), ii) background noise quality (BAK), and iii) the overall quality (OVRL) of the audio. The developed metric is highly correlated with human ratings, with a Pearson's Correlation Coefficient (PCC)=0.94 for SIG and PCC=0.98 for BAK and OVRL. This is the first non-intrusive P.835 predictor we are aware of. DNSMOS P.835 is made publicly available as an Azure service.

Index Terms— Speech, Perceptual Speech Quality, Objective Metric, Deep Noise Suppressor, Metric, P.835.

1. INTRODUCTION

Subjective evaluation of speech quality is the most reliable way to evaluate Speech Enhancement (SE) methods [3]. However, subjective tests are not easily scalable as they require a considerable number of listeners, the process is laborious, time-consuming, and expensive. Conventional objective speech quality metrics such as Perceptual Evaluation of Speech Quality (PESQ) [4], Perceptual Objective Listening Quality Analysis (POLQA) [5], VisQOL [6] and Signal to Distortion Ratio (SDR) are widely used to evaluate Speech Enhancement (SE) algorithms optimized for human perception. Some of these metrics are designed to predict the subjective Mean Opinion Score (MOS) obtained using the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) Recommendation P.800 [7]. However, they are shown to correlate poorly with human rating when used for SE tasks that involve perceptually invariant transformations [3]. Also, intrusive metrics cannot be used to evaluate real recordings when a clean reference is unavailable in realistic scenarios.

2. RELATED WORK

The subjective test ITU-T P.835 [2] provides the speech quality (SIG), background noise quality (BAK), and overall quality (OVRL). Hu and Loizou [8] showed an accurate linear model of OVRL can be estimated as a function of SIG and BAK. Naderi and Cutler [9] used this linear relationship to analyze the results of the 3rd Deep Noise Suppression challenge [10] to estimate the potential improvement in OVRL given a noise suppressor that maximized BAK. Hu and Loizou [8] released an intrusive speech quality assessment tool based on P.835, with a correlation to subjective quality of PCC(SIG)=0.70, PCC(CBAK)=0.58, PCC(OVRL)=0.73 using a synthetic training and test set. A commercial tool, 3QUEST [11], is used to measure the speech (S-MOS), noise (N-MOS), and overall (G-MOS) quality of speech as part of the ETSI EG 202 396-3 standard for mobile telephone quality. This intrusive model has good performance, PCC(S-MOS)=0.92, PCC(N-MOS)=0.94, PCC(G-MOS)=0.94 by condition; it was trained with 179 conditions and tested with 81 conditions, but the duration of training data and testing data is not reported [11].

ITU-T Recommendation P.563 is a non-intrusive technique and can directly operate on the degraded signal [12]. However, it was developed for narrow-band applications, works on limited impairment types, but correlates poorly with human ratings [13]. Recently, Deep Neural Networks (DNNs) based approaches have been proposed to estimate the speech quality scores [14, 15, 13, 16, 17, 18, 19, 20]. Some of these learning-based approaches use other objective metrics as the ground truth to train their speech quality predictor. Other methods use MOS obtained using P.800 as the ground truth to train their models. In [21], the authors trained the model to identify the Just Noticeable Difference (JND). MOS predictors trained on actual human ratings are more reliable than the ones trained to predict other objective metrics like PESQ or POLQA. The accuracy and robustness of the learned models depend on the quality of the human labels and also

the quantity and diversity of the audio clips. A comparison of some common DNN-based non-intrusive speech quality assessment (NI-SQA) methods is given in Table 1. ACR is Absolute Catagorgy Rating [7]. DNSMOS P.835 is the first P.835 based NI-SQA model we are aware of.

In [16], we show that the NI-SQA metric called DNSMOS trained using subjective quality labels is more robust and reliable than some of the other popular intrusive metrics. DNSMOS is used to do model training and model selection during noise suppression development. DNSMOS is also used for doing ablation studies for noise suppressors [22, 23]. DNSMOS has been quite popular, with over a hundred researchers using it after several months of releasing it.

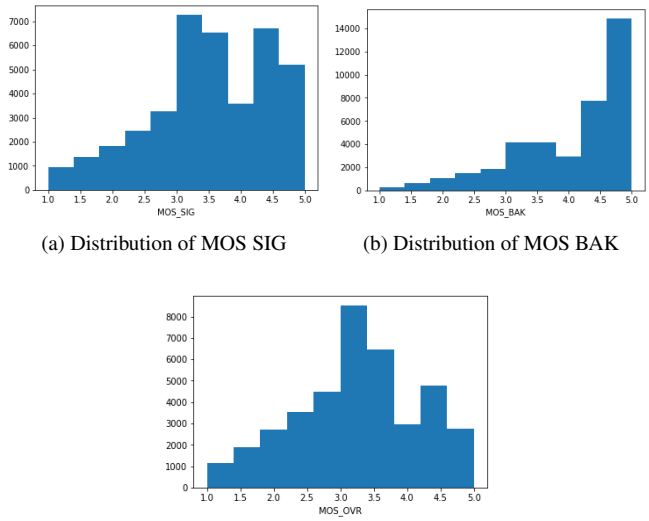
However, DNSMOS only gives the overall score of the audio clip. In this paper, we extend that work to predict the quality of speech (SIG), background noise (BAK), and overall quality (OVRL) of the audio clip. We use the subjective quality labels obtained from ITU-T P.835 from Deep Noise Suppression (DNS) Challenge 3 [10] and the noisy clips processed by several noise suppression models internally at Microsoft. The labels were obtained using our crowdsourcing-based extension of P.835 described in [9]. The model uses log power spectrogram as input features to a Convolutional Neural Network (CNN) based model. It can be used to stack rank different DNS methods based on MOS estimates with great accuracy and hence the name DNSMOS P.835. We are providing DNSMOS P.835 as an Azure service for other researchers to use. The details of the API are at `www.microsoft.com/en-us/research/dns-challenge/dnsmos`.

3. DATA AND SUBJECTIVE RATINGS

We used the labeled data from the DNS Challenge V3 [10] to train DNSMOS P.835. The DNS Challenge V3 test set comprised of 600 noisy speech clips processed by about 40 different noise suppression models. The real recordings in the test set were captured in a variety of noise types and Signal to Noise Ratio (SNR) and target levels. The test set is comprised of over 100 noise types and speakers. More details about the creation of these test sets can be found in [10]. The speech quality ratings of the processed clips varied from very poor (MOS=1) to excellent (MOS=5) for SIG, BAK, and OVRL. The distribution of the MOS scores in the training data is shown in Figure 1. The scores are highly skewed with most ratings populated in the range 3<MOS<4 and fewer ratings in both the tails for SIG and OVRL. However, BAK is highly skewed towards MOS > 4.

Table 1: Comparison of some DNN NI-SQA methods

Model	Data size (hours)	Data type
[20]	5.2	ACR
WAWENETS [19]	17	ACR
[13]	27.7	ACR
[15]	27.7	ACR
SESQA [24]	45.2	ACR, JND
DNSMOS [16]	300	ACR
DNSMOS P.835	75	P.835 ACR



(a) Distribution of MOS SIG

(b) Distribution of MOS BAK

(c) Distribution of MOS OVRL

Fig. 1: Distribution of the training set

The subjective quality ratings are obtained in several P.835 runs conducted over several months. Multiple noise suppression methods are compared in each P.835 run. Each P.835 run included the best-performing noise suppressor, original noisy speech, and a couple of methods with intermediate perceptual quality from previous runs as anchors. Hence, some of the clips were rated multiple times. In total, we have about 30,000 audio clips with associated MOS scores as ground truth. The average length of each audio clip was about 9 seconds, giving us a total of 75 hours of data.

A subset of the dataset is summarized in Figure 2. What makes this dataset unique is (1) it is by far the largest P.835 dataset we know of and the only one used to train a DNN non-intrusive speech quality assessment model, and (2) the 40 deep noise suppression models used in the dataset gives a large variety of suppression artifacts we think is needed to generalize a speech quality assessment model for noise suppressors.

4. DNSMOS P.835

4.1. Features

Recently, researchers have seen success in learning features within the model for tasks such as SE [25], speech, and music

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
38	(0.00)	0.05	(0.07)	0.12	0.03	(0.02)	0.03	0.04
36	0.01	0.07	(0.00)	0.16	(0.17)	(0.11)	0.01	0.04
Noisy	0 (4.02)	0 (3.83)	0 (3.93)	0 (3.8)	0 (3.97)	0 (3.87)	0 (3.89)	0.04
40	(0.04)	(0.08)	(0.17)	0.03	(0.21)	(0.23)	(0.10)	0.04
33	(0.13)	(0.15)	(0.21)	0.03	(0.13)	(0.23)	(0.12)	0.04
13	(0.15)	(0.15)	(0.17)	0.04	(0.26)	(0.24)	(0.13)	0.04
19	(0.21)	(0.18)	(0.13)	0.06	(0.26)	(0.31)	(0.15)	0.04
34	(0.16)	(0.17)	(0.12)	0.08	(0.41)	(0.36)	(0.17)	0.04
1	(0.09)	(0.24)	(0.20)	0.06	(0.38)	(0.34)	(0.17)	0.04
18	(0.35)	(0.48)	(0.47)	(0.06)	(0.73)	(0.54)	(0.39)	0.04
30	(0.44)	(0.80)	(0.23)	(0.24)	(0.53)	(0.49)	(0.43)	0.05
20	(0.48)	(0.65)	(0.38)	(0.18)	(0.62)	(0.60)	(0.45)	0.04
8	(0.47)	(0.57)	(0.37)	(0.17)	(0.97)	(0.76)	(0.52)	0.05
31	(0.45)	(0.63)	(0.54)	(0.26)	(0.84)	(0.68)	(0.53)	0.05
Baseline	(0.49)	(0.68)	(0.49)	(0.27)	(0.74)	(0.72)	(0.54)	0.04
4	(0.62)	(0.69)	(0.65)	(0.51)	(0.50)	(0.73)	(0.61)	0.05
22	(0.46)	(0.94)	(0.63)	(0.33)	(0.85)	(0.80)	(0.62)	0.05
12	(0.54)	(1.08)	(0.56)	(0.31)	(1.16)	(0.86)	(0.69)	0.05
11	(0.82)	(1.09)	(0.71)	(0.41)	(1.03)	(0.84)	(0.76)	0.05
37	(0.72)	(0.99)	(0.64)	(0.43)	(1.10)	(1.01)	(0.78)	0.05
28	(0.94)	(1.39)	(0.87)	(0.66)	(1.55)	(1.12)	(1.03)	0.05

(a) Speech MOS

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
36	1.92	2.73	1.82	1.54	2.39	2.31	2.05	0.02
1	1.90	2.59	1.71	1.54	2.25	2.21	1.98	0.03
18	1.82	2.63	1.61	1.42	2.07	2.24	1.91	0.03
33	1.78	2.42	1.54	1.40	2.27	2.13	1.87	0.03
13	1.73	2.28	1.48	1.34	2.01	1.94	1.74	0.03
22	1.77	2.20	1.37	1.36	1.91	1.97	1.73	0.03
34	1.78	2.15	1.51	1.28	1.84	1.84	1.68	0.03
8	1.62	2.01	1.36	1.17	1.81	1.83	1.59	0.03
37	1.69	2.14	1.46	1.13	1.61	1.78	1.58	0.04
19	1.48	1.92	1.36	1.14	1.84	1.68	1.52	0.03
12	1.82	2.13	1.44	1.07	1.13	1.58	1.47	0.04
Baseline	1.35	1.68	1.32	0.92	0.97	1.64	1.28	0.04
20	1.55	1.61	1.29	1.06	1.04	1.34	1.28	0.04
11	1.39	1.52	0.95	0.86	1.43	1.30	1.20	0.04
40	1.50	1.52	0.97	0.86	1.10	1.21	1.15	0.04
31	1.24	1.70	1.08	0.72	1.21	1.21	1.12	0.04
28	1.01	1.34	0.91	0.78	0.80	1.22	1.00	0.04
30	1.53	1.23	0.87	0.64	0.62	0.59	0.85	0.05
4	0.24	0.49	0.26	0.13	0.33	0.16	0.23	0.04
Noisy	0 (2.86)	0 (1.93)	0 (2.91)	0 (3.11)	0 (2.14)	0 (2.3)	0 (2.6)	0.04
38	(0.09)	(0.04)	(0.04)	(0.03)	0.04	0.01	(0.02)	0.04

(b) Background Noise MOS

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
36	0.89	1.51	0.79	0.80	1.14	1.11	1.01	0.04
1	0.85	1.16	0.65	0.69	0.89	0.92	0.85	0.04
33	0.74	1.16	0.56	0.57	1.06	0.93	0.81	0.04
13	0.76	1.14	0.60	0.60	0.97	0.90	0.80	0.04
34	0.69	1.13	0.59	0.64	0.76	0.75	0.74	0.04
19	0.61	0.98	0.58	0.57	0.90	0.74	0.71	0.04
18	0.59	1.02	0.40	0.55	0.55	0.75	0.64	0.04
40	0.79	0.94	0.37	0.40	0.62	0.63	0.60	0.04
8	0.41	0.79	0.37	0.39	0.25	0.41	0.42	0.04
22	0.51	0.50	0.18	0.30	0.42	0.42	0.39	0.05
20	0.44	0.48	0.28	0.36	0.29	0.39	0.38	0.04
31	0.37	0.60	0.20	0.20	0.28	0.35	0.32	0.04
Baseline	0.25	0.47	0.31	0.21	0.21	0.41	0.30	0.04
12	0.39	0.33	0.29	0.23	(0.00)	0.28	0.25	0.04
30	0.44	0.27	0.31	0.12	0.16	0.17	0.22	0.04
37	0.18	0.41	0.15	0.13	0.13	0.20	0.19	0.04
11	0.07	0.25	(0.08)	0.09	0.16	0.25	0.14	0.04
38	(0.12)	0.04	(0.10)	0.02	0.09	0.06	0.01	0.04
Noisy	0 (3.03)	0 (2.28)	0 (3)	0 (3.04)	0 (2.57)	0 (2.52)	0 (2.77)	0.04
28	(0.12)	(0.07)	(0.10)	(0.12)	(0.44)	(0.02)	(0.13)	0.04
4	(0.22)	0.13	(0.26)	(0.27)	0.02	(0.15)	(0.15)	0.04

(c) Overall MOS

Fig. 2: Track 1 results for the 3rd Deep Noise Suppression Challenge

synthesis [26] and to learn acoustic models [27]. They show that using the time-domain waveform requires a larger model trained on a larger and diverse data set to ensure generalization. The ground truth MOS scores are obtained for audio clips with an average length of 9 seconds sampled at 16 kHz. This leads to a very large input dimension if we are treating it as a vector and the model will require many layers to compress and extract input features. Instead, we used log power spectrogram as input feature extracted over 9 seconds duration as it correlates well with human perception and is proven to work very well for analyzing speech quality [15]. For spectral features, we used a frame size of 20 ms with a Hamming window and a hop length of 10 ms. The input features are then converted to dB scale.

4.2. Prediction model

For predicting the MOS scores, we explored different configurations of CNN-based models. The architecture for the best performing model is shown in Table 2. The input to the model is log power spectrogram with a 320 FFT size computed over a clip of length 9 seconds sampled at 16 kHz with a frame size of 20 ms and hop length of 10 ms. This results in an input dimension of 900 x 161. We trained two different models with almost the same architecture except for the last layer. One model is trained to predict all 3 outputs (SIG, BAK, OVRL) and the other model is trained to predict only SIG. The reason is we found the prediction of SIG is a much harder task and is less correlated with BAK and OVRL. The models were trained with a batch size of 32 using the Adam optimizer and MSE loss function until the loss saturated. We experimented by adding batch normalization layers after every Conv layer in Table 1. However, adding batch normalization reduces the prediction accuracy of low volume clips. Humans tend to give lower ratings to clips with low amplitudes [28]. We want the model to capture the variations in the target levels of the data. Hence, we avoid any kind of feature normalization. We also explored different network architectures including CNN followed by LSTM. The model in Table 2 generalized the best and was of least complexity.

5. EXPERIMENTAL RESULTS

5.1. Test set

The unseen real test set used to validate the trained model consists of P.835 evaluation of 17 different Microsoft internal noise suppression models on an unseen set of 850 clips. The clips span various categories like emotional, English, Non-English with and without tonal languages, and stationary noises. This unseen test set was created for a future DNS challenge and has similar categories as the training data, adding mouse clicks and improving the quality of emotional speech. The test set was created with crowdsourcing using the method described in [10].

Table 2: DNSMOS P.835 Prediction Model

Layer	Output dimension
Input	900 x 120 x 1
Conv: 128, (3 x 3), 'ReLU'	900 x 161 x 128
Conv: 64, (3 x 3), 'ReLU'	900 x 161 x 64
Conv: 64, (3 x 3), 'ReLU'	900 x 161 x 64
Conv: 32, (3 x 3), 'ReLU'	900 x 161 x 32
MaxPool: (2 x 2), Dropout(0.3)	450 x 80 x 32
Conv: 32, (3 x 3), 'ReLU'	450 x 80 x 32
MaxPool: (2 x 2), Dropout(0.3)	225 x 40 x 32
Conv: 32, (3 x 3), 'ReLU'	112 x 20 x 32
MaxPool: (2 x 2), Dropout(0.3)	112 x 15 x 32
Conv: 64, (3 x 3), 'ReLU'	112 x 20 x 64
GlobalMaxPool	1 x 64
Dense: 128, 'ReLU'	1 x 128
Dense: 64, 'ReLU'	1 x 64
Dense: 1 or 3	1 x 1 or 1 x 3

5.2. Evaluation metric

PCC or MSE between the predictions of the developed objective metric and the ground truth human ratings is commonly used to measure the accuracy of the model [15, 16]. From [9], we know that P.835 is highly repeatable between runs when averaged across a set of clips per condition, which can be formed by grouping clips enhanced by a particular SE model or based on other criteria like SNR or reverb RT60 times. The PCC computed on the average of ratings per group across different runs is $>$0.9. We also found that PCC computed on the same clips but from two different P.835 runs is only about 0.7-0.8 due to the high rating noise per clip.

Hence, for stack ranking different noise suppressors we evaluate by computing the average of ratings across the entire test set for each model. Therefore, we compute Spearman's Rank Correlation Coefficient (SRCC) and PCC between averaged human ratings and averaged DNSMOS per model. SRCC gives us the stack ranking accuracy of various SE models.

Table 3: Model and clip level correlation of DNSMOS P.835 with human ratings

Type	SIG	BAK	OVRL
Model PCC	0.94	0.98	0.98
Model SRCC	0.95	0.99	0.98
Clip PCC	0.71	0.83	0.82
Clip SRCC	0.72	0.82	0.81

5.3. Results

Table 3 shows the per model and per clip PCC and SRCC between human ratings and DNSMOS P.835 on the unseen test set described in Section 5.1. When DNSMOS is aggregated by model the results are excellent, though it still shows an area for improvement in SIG. The results on this unseen test set show DNSMOS P.835 generalizes well, at least for these categories of noises and environments. We can not compare DNSMOS P.835 with other metrics since it is the first NI-SQA metric for P.835 we are aware of.

The clip level correlation of two noise suppression models on the same dataset but using N=30 ratings per clip instead of N=5 used in the per model correlation to give us better accuracy for the human ratings. These results show DNSMOS P.835 has good per clip performance also, though of course not as good as when aggregated at the model level.

6. CONCLUSION AND FUTURE WORK

DNSMOS P.835 is an accurate speech quality metric designed to stack rank noise suppressors with great accuracy. We attribute the excellent performance of DNSMOS P.835 to (1) a large high-quality dataset, (2) a limited speech quality impairment category, (3) significant optimizations on the model architecture and training, and (4) aggregation by noise suppression model. The per clip performance can be improved by significantly increasing the number of ratings per clip, which is currently only 5 because of cost restrictions. We can also expand the complexity of the model to further improve performance.

7. REFERENCES

[1] ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*, International Telecommunication Union, Geneva, 2018.

[2] ITU-T Recommendation P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, International Telecommunication Union, Geneva, 2003.

[3] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. INTERSPEECH 2019*, pp. 1816–1820, 2019.

[4] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, vol. 2, pp. 749–752 vol.2.

[5] John Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II-Perceptual Model," *AES: Journal of the Audio Engineering Society*, vol. 61, pp. 385–402, 06 2013.

[6] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte, "ViSQOL: an objective speech quality model," *Eurasip Journal on Audio, Speech, and Music Processing*, , no. 1, pp. 13, 2015.

[7] "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," Feb 1998.

[8] Yi Hu and Philipos C Loizou, "Evaluation of objective measures for speech enhancement," in *International Conference on Spoken Language Processing*, 2006.

[9] Babak Naderi and Ross Cutler, "Subjective evaluation of noise suppression algorithms in crowdsourcing," in *INTERSPEECH*, 2021.

[10] Chandan K A Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *INTERSPEECH*, 2021.

[11] Head Acoustics Application Note, "3QUEST: 3-fold Quality Evaluation of Speech in Telecommunications Systems," 2008.

[12] "ITU-T Recommendation P.563: Single-ended method for objective speech quality assessment in narrowband telephony applications," 2004.

[13] 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, pp. 631–635.

[14] Xuan Dong and Donald S Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *ICASSP*. IEEE, 2020, pp. 911–915.

[15] Hannes Gamper, Chandan KA Reddy, Ross Cutler, Ivan J Tashev, and Johannes Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *WASPAA*. IEEE, 2019, pp. 85–89.

[16] Chandan K A Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2021, pp. 6493–6497.

[17] Jasper Ooster and Bernd T Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *ICASSP*. IEEE, 2019, pp. 636–640.

[18] 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 *INTERSPEECH*, 2018.

[19] Andrew A Catellier and Stephen D Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *ICASSP*. IEEE, 2020, pp. 331–335.

[20] Benjamin Cauchi, Kai Siedenburg, Joao F Santos, Tiago H Falk, Simon Doclo, and Stefan 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.

[21] Pranay Manocha, Adam Finkelstein, Zeyu Jin, Nicholas J Bryan, Richard Zhang, and Gautham J Mysore, "A differentiable perceptual audio metric learned from just noticeable differences," in *INTERSPEECH*, 2020.

[22] Shubo Lv, Yanxin Hu, Shimin Zhang, and Lei Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," in *INTERSPEECH*, 2021.

[23] Andong Li, Wenzhe Liu, Xiaoxue Luo, Guochen Yu, Chengshi Zheng, and Xiaodong Li, "A simultaneous denoising and dereverberation framework with target decoupling," in *INTERSPEECH*, 2021.

[24] Joan Serrà, Jordi Pons, and Santiago Pascual, "Sesqa: semi-supervised learning for speech quality assessment," in *ICASSP*. IEEE, 2021, pp. 381–385.

[25] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," in *INTERSPEECH*, 2020.

[26] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo, "WaveCycleGAN2: Time-domain neural post-filter for speech waveform generation," *arXiv preprint arXiv:1904.02892*, 2019.

[27] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *ICASSP*, 2015, pp. 4624–4628.

[28] Côté Nicolas, Valérie Gautier-Turbin, and Sebastian Möller, "Influence of loudness level on the overall quality of transmitted speech," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.