Pattern Recognition Letters 166 (2023) 159–163

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Differentiable Mean Opinion Score Regularization for Perceptual Speech Enhancement

Tomer Rosenbaum [a,*], Israel Cohen [a], Emil Winebrand [b], Ofri Gabso [b]

[a] Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel
[b] Insoundz Ltd., Tel Aviv, Israel

ARTICLE INFO

Article history:
Received 29 April 2022
Revised 6 January 2023
Accepted 16 January 2023
Available online 19 January 2023

Edited by Maria De Marsico

Keywords:
Speech enhancement
Mean opinion score
Speech quality assessment
Speech naturalness assessment

ABSTRACT

Many speech enhancement methods require perceptual quality metrics for evaluation. The "holy grail" of perceptual speech quality assessment is human subjective ratings, known as the mean opinion score. However, acquiring human ratings is time-consuming, laborious, and expensive. Existing objective quality metrics, on the other hand, are efficient and easy to compute but do not correlate well with human ratings. In this paper, we propose a relatively lightweight deep-learning-based model to predict the human ratings of speech signals. Since it is differentiable, it can be easily employed as a perceptual regularization to improve existing deep-learning-based speech enhancement methods. Experimental results demonstrate that the predictions of our proposed model correlate well with human judgments. We present application in speech enhancement and show that, interestingly, while there is a degradation in performance in terms of traditional objective metrics, there is a significant improvement in the perceptual quality and the naturalness of the enhanced speech.

1. Introduction

Quantification and assessment of speech quality are crucial for objective and subjective evaluation of speech enhancement methods. The "gold standard" of speech quality assessment in terms of human perception is the mean opinion score (MOS) of human subjective ratings. The process of performing subjective tests requires a large number of human listeners and can be expensive and time-consuming, and therefore is not scalable. Traditional objective metrics, on the other hand, such as Perceptual Evaluation of Speech Quality (PESQ) [1], Perceptual Objective Listening Quality Analysis (POLQA) [2], Virtual Speech Quality Objective Listener (ViSQOL) [3], and Short-Time Objective Intelligibility (STOI) [4], can be computed efficiently and are interpretable. However, the main drawback of such metrics is that they usually correlate poorly with subjective human MOS ratings and are sensitive to perceptually-invariant transformations [5,6]. Secondly, those metrics are not differentiable, and hence deep-learning-based speech enhancement models cannot be optimized with respect to such metrics using gradient-based optimization methods.

Recent research has focused on designing and developing new metrics to overcome those two issues. In the context of non-differentiability, Perceptual Metric for Speech Quality Evaluation (PMSQE) [7] was proposed as a differentiable approximation of PESQ. Fu et al. [8] proposed to model PESQ with a generative adversarial network (GAN). Both methods can be employed in any deep-learning-based training framework. However, it seems that they fail to generalize well to unseen perturbations. Recently, Xu et al. [9] proposed a novel non-intrusive PESQNet deep neural network to estimate the PESQ scores of enhanced speech signals without knowing the corresponding clean reference speech. PESQNet shows impressive correlation with ground-truth PESQ scores. In addition, this model can be employed to fine-tune a deep-learning-based noise suppressor and improve its performance in terms of PESQ.

In the context of the correlation of speech quality metrics with human MOS ratings, recent works have proposed to train a deep learning model to predict the MOS ratings from speech input, based on supervised speech data with corresponding MOS ratings. Deep Noise Suppression MOS (DNSMOS) [10] is a reference-free deep learning model that is effective for the evaluation of noise suppressors. DNSMOS is trained on speech outputs of various noise suppressors with corresponding human MOS ratings. The trained model predicts MOS in terms of Absolute Category Rating (ACR), i.e., a score on a scale from 1 (very poor) to 5 (excellent). MOSNet [11] is trained to assess the quality of voice

* Corresponding author.
E-mail addresses: tomer11r@campus.technion.ac.il (T. Rosenbaum), icohen@ee.technion.ac.il (I. Cohen), emil.winebrand@insoundz.com (E. Winebrand), ofri@insoundz.com (O. Gabso).

https://doi.org/10.1016/j.patrec.2023.01.011
0167-8655/

conversion algorithms. Manocha et al. [12,13] proposed a reference-based model trained on the Just-Noticeable Differences (JND) between the speech signal and the corresponding reference clean signal. The JND labels are obtained using an active learning approach. After training, the model can be interpreted as a distance function sensitive to JND between two speech signals in terms of human perception. Semi-supervised learning for Speech Quality Assessment (SESQA) [14] is trained to minimize an objective function that includes subjective, objective, and JND measures and predicts MOS in terms of ACR. Recently, Manocha et al. [15] proposed an interesting framework for Non-matching Reference-based Speech Quality Assessment (NORESQA) to measure distance (in terms of speech quality) between two signals that are not necessarily matching (in terms of speech content and speaker). It is based on the concept that humans can easily compare the perceptual quality of two speech signals even if they are not matching. NORESQA is trained without human MOS ratings and is surprisingly competitive with data-driven methods such as DNSMOS.

This paper proposes a reference-free model that can be easily embedded in training frameworks for deep-learning-based speech processing models. This can be interpreted as a perceptual regularization term that enforces the outputs of such models to have high perceptual quality in terms of human MOS ratings. First, we collect data from speech clips and corresponding human MOS ratings by performing online subjective tests according to the International Telecommunication Union (ITU-T) Recommendation P.808 standard [16]. This recommendation specifies the experiment design, test procedure, and data analysis for human subjective MOS in terms of ACR, and is considered a reliable approach to crowdsourcing speech quality assessment [17]. We train our proposed model to predict frame-wise absolute speech quality ratings using the collected data. We show that the predictions of the trained model correlate well with human MOS ratings. To show how the model can be leveraged, we present an application in deep-learning-based speech enhancement. We conduct experiments to validate the effectiveness of the proposed application and offer an interesting observation. While our proposed perceptual regularization clearly improves the performance of an existing speech enhancement deep learning model in terms of human judgments and speech naturalness, there is a degradation in performance in terms of traditional objective measures such as PESQ and STOI. The observations suggest that PESQ and STOI are more sensitive to degradation in speech that might no be perceived by the human ear. Code, model weights and audio samples will be made available upon publication.[1]

The remainder of this paper is organized as follows: In Section 2, we present the MOS data collection framework. Section 3 describes the proposed model. Experiments and performance evaluation are discussed in Section 4. Finally, conclusions are presented in Section 5.

2. MOS Data Acquisition

We acquire a dataset of speech recordings in different conditions with corresponding human MOS rating labels to train and evaluate the MOS predictor. More specifically, we collect $N_{\text{train}} = 600$ clean speech recordings for train and $N_{\text{test}} = 130$ clean recordings for test from the Deep Noise Suppression (DNS) challenge dataset [18]. The minimal length of every speech clip is 10 seconds. Each clean recording is convolved with room impulse responses (RIRs), corresponding to microphone arrays in different rooms. Background noise is added to the signal such that the signal-to-noise ratio (SNR) varies from -1 dB to 24 dB. The resulting multichannel signal, consisting of reverberation and background noise, is fed to a weighted prediction error (WPE) model [19] to reduce reverberation and then fed to the minimum variance distortionless response (MVDR) beamformer [20] to produce an enhanced single-channel speech signal.

We denote the training and test sets of the enhanced clips as $\{x_i\}_{i=1}^{N_{\text{train}}}$ and $\{\tilde{x}_i\}_{i=1}^{N_{\text{test}}}$, respectively. To get the MOS label, we conduct ITU-T P.808 tests [16] using Amazon Mechanical Turk (AMT) crowdsourcing platform, where the perceptual quality of each enhanced speech clip (in terms of ACR) is rated by $N_r = 20$ different users. We denote the MOS rating of the i-th speech clip by the j-th user as r_i^j if the clip is from the trainset and $\tilde{r}_i^j$ if the clip is from the testset. In the testset, the average MOS of the clip i over the 20 users is denoted as $\tilde{R}_i$ and it is considered as the MOS rating of clip i. To summarize, the trainset $\left\{\{x_i, r_i^j\}_{j=1}^{N_r}\right\}_{i=1}^{N_{\text{train}}}$ consists of $N_{\text{train}} \cdot N_r = 12000$ labeled samples and the testset $\left\{\tilde{x}_i, \tilde{R}_i\right\}_{i=1}^{N_{\text{test}}}$ consists of $N_{\text{test}} = 130$ labeled samples.

3. Proposed Model

Let $x(t)$ be a waveform speech signal in the discrete-time domain consisting of τ samples. Denote the short-time Fourier transform (STFT) magnitude of $x(t)$ as $X = |\text{STFT}(x)| \in \mathbb{R}^{T_\tau \times F}$ where T_τ and F are the number of STFT time and frequency bins, respectively. We employ the architecture of MOSNet [11] as a differentiable deep learning-based human MOS predictor with trainable parameters ϕ, denoted as $\mathbf{M}_\phi$. The model is shown in Fig. 1. Given an utterance $x(t)$, the model $\mathbf{M}_\phi$ is fed with X, followed by a series of four 2D convolution blocks, a bidirectional long short-term memory (BLSTM) layer, and two frame-wise fully connected (FC) layers. The output $\mathbf{M}_\phi(X) \in [1,5]^{T_\tau}$ consists of T_τ frame-wise MOS estimations. The overall predicted human MOS rating of the utterance $x(t)$ is considered the mean of $\mathbf{M}_\phi(X)$.

3.1. Training

For training, we use the trainset described in Section 2. Given pairs of speech clips x and their corresponding MOS labels r, the trainable parameters ϕ are optimized to minimize the mean-square-error (MSE) objective function between the network's predictions and the acquired MOS rating:

$$\mathcal{L}_{\text{MSE}}(\mathbf{M}_\phi(X), r) = \mathbb{E}\left[\left\|\mathbf{M}_\phi(X) - r\right\|_2^2\right] \tag{1}$$

where $\mathbb{E}[\cdot]$ stands for expectation over the distribution of the trainset and $\|\cdot\|_p$ stands for the L_p norm. To improve generalization, we employ drop-out units with parameter $p = 0.3$ after the BLSTM layer and after the first FC layer. This loss term enforces the overall predicted rating (i.e., the mean of $\mathbf{M}_\phi(X)$) to be close to the MOS rating. We train until convergence and denote the optimized parameters as ϕ^*. For convenience, we denote the trained model as $\mathbf{M} \triangleq \mathbf{M}_{\phi^*}$.

3.2. Application in Speech Enhancement

As shown in Section 4, the trained model $\mathbf{M}$ correlates well with human MOS. Hence, it can be leveraged as a prior that improves the performance of existing deep-learning-based speech enhancement models in terms of human perception. Let $y(t)$ and $x(t)$ be a corrupted speech signal and the corresponding ground-truth reference clean speech signal, respectively. Let f_θ be any deep-learning-based speech enhancement model with trainable weights θ, i.e., the model f_θ is fed with $y(t)$ and trained to produce an enhanced signal $\hat{x}_\theta(t)$ such that $f_\theta(y(t)) = \hat{x}_\theta(t) \approx x(t)$. Since $\mathbf{M}$

[1] https://github.com/tomermistrix/mosnet-speech-enhancement.

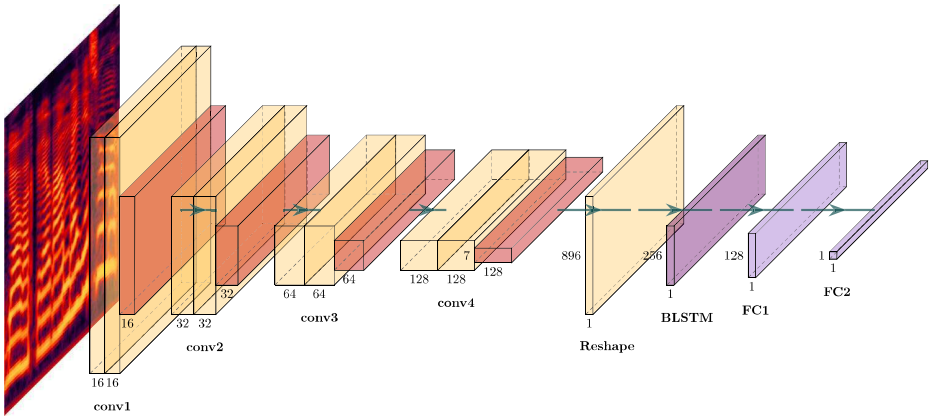

Fig. 1. MOSNet architecture. The input STFT magnitude is fed into a series of 2D convolutional blocks, a BLSTM layer, and two frame-wise FC layers. The output is frame-wise human MOS prediction.

Table 1
Correlation in terms of PCC and SRCC of predicted MOS of our model (MOSNet), MOS, and predicted MOS of DNSMOS.

PCC	MOSNet	MOS	DNSMOS
MOSNet	1	**0.909**	0.853
MOS	**0.909**	1	0.797
DNSMOS	0.853	0.797	1
SRCC	**MOSNet**	**MOS**	**DNSMOS**
MOSNet	1	**0.903**	0.844
MOS	**0.903**	1	0.775
DNSMOS	0.844	0.775	1

is implemented as a differentiable neural network, we propose an objective term that can be employed in the training framework of f_θ. Note that in this setting, the trainable weights of $\mathbf{M}$ (i.e., ϕ^*) are fixed during the training of f_θ. Compared to previous works, such as [12,13], where the proposed objective terms require the ground-truth clean speech (and hence suitable for supervised learning), our proposed objective term does not require a reference ground-truth clean signal. It hence can be seen as a perceptual regularization constraint on the output that can be employed in any learning framework (i.e., both supervised and unsupervised settings). It is defined as:

$$\mathcal{L}_{\text{MOS}}(\hat{x}_\theta) = \mathbb{E}\big[5 - \text{Mean}\big(\mathbf{M}(\hat{X}_\theta)\big)\big]. \tag{2}$$

The term can be added to any existing loss function. Minimization of (2) with respect to θ enforces the model outputs to have a large MOS rating.

4. Experimental Results

4.1. Correlation with Human MOS

To show that the trained model predicts MOS ratings that correlate well with human ratings, we predict MOS ratings for the test set from Section 2. For comparison, we also predict ratings using the publicly available DNSMOS [10]. The results in the Pearson Correlation Coefficient (PCC) [21] and Spearman's Rank Correlation Coefficient (SRCC) are shown in Table 1. As can be seen, our proposed model outperforms DNSMOS in terms of correlation with human MOS. Scatter plots of the predicted ratings and DNSMOS ratings compared to the MOS ratings are presented in Figs. 2(b) and (c), respectively. For further visualization, we present in Fig. 2(a) the sample-wise predicted MOS, MOS, and DNSMOS ratings. For convenience, the samples are sorted in ascending order with respect to the MOS ratings. The observations in Fig. 2 indicate that while the predictions of both our proposed model and DNSMOS correlate well with human ratings when the human MOS rating is relatively large, our model is more accurate when the human MOS rating is small.

4.2. Speech Enhancement

To demonstrate the effectiveness of the proposed approach, we consider the speech enhancement model from [22] as f_θ. This model, which is based on the Demucs architecture [23], is known as an effective model for speech denoising. Given a trainset consisting of N noisy speech signals $\{y_i\}_{i=1}^N$ and corresponding ground-truth clean signals $\{x_i\}_{i=1}^N$, we train the model to enhance noisy signals using our proposed perceptual regularization. We use the official implementation of the model architecture in [22] and initialize the model weights using the publicly available "master" pre-trained model weights. The initialized "master" model was trained using data from both Valentini [24] and DNS challenge [18].

4.2.1. Training

For training and evaluation, we use the DNS challenge [18] benchmark. More specifically, we generate a trainset consisting of pairs of noisy and corresponding ground-truth clean 5-second speech signals such that the SNR level of the pairs is uniformly distributed in the range $[-1, 24]$ dB. For comparison, we consider two models:

- *Baseline:* we fine-tune the "master" model on our generated trainset until convergence to minimize the loss function:

$$\mathcal{L}_{\text{base}}(x, \hat{x}_\theta) = \mathbb{E}\big[\big\|x - \hat{x}_\theta\big\|_1\big] + \lambda_{\text{STFT}}\mathcal{L}_{\text{STFT}}(x, \hat{x}_\theta) \tag{3}$$

  where x and $\hat{x}_\theta$ are the ground-truth speech signals and the predicted enhanced outputs, respectively, $\mathcal{L}_{\text{STFT}}$ is the multi-resolution STFT loss proposed in [22], and $\lambda_{\text{STFT}} = 0.1$ is a weight hyper-parameter.
- *MOSNet loss:* here, we take into account the proposed approach and fine-tune the "master" model until convergence to minimize:

$$\mathcal{L}_{\text{MOSNet}}(x, \hat{x}_\theta) = \mathcal{L}_{\text{base}}(x, \hat{x}_\theta) + \lambda_{\text{MOS}}\mathcal{L}_{\text{MOS}}(\hat{x}_\theta) \tag{4}$$

  where λ_{MOS} is a weight hyper-parameter and $\mathcal{L}_{\text{MOS}}$ is the term presented in (2). Note that the term $\mathcal{L}_{\text{MOS}}$ produces values on the range [0,4]. When fine-tuning the baseline model, we observed that $\mathcal{L}_{\text{base}}$ typically produces small values on the range [0.01,0.05]. Based on this observation and in order to set both terms in (4) significant, we set $\lambda_{\text{MOS}} = 0.01$.

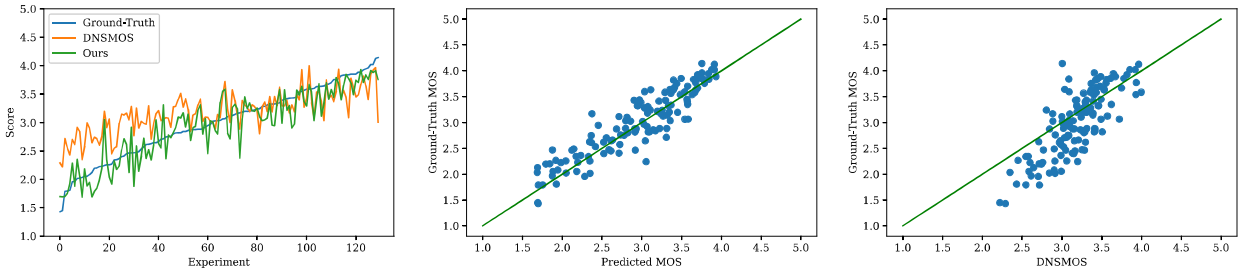

Fig. 2. (a) Ours predicted MOS ratings vs. DNSMOS predicted ratings, compared to MOS, (b) scatter plot of our predictions vs. MOS ratings, (c) scatter plot of DNSMOS predictions vs. MOS ratings.

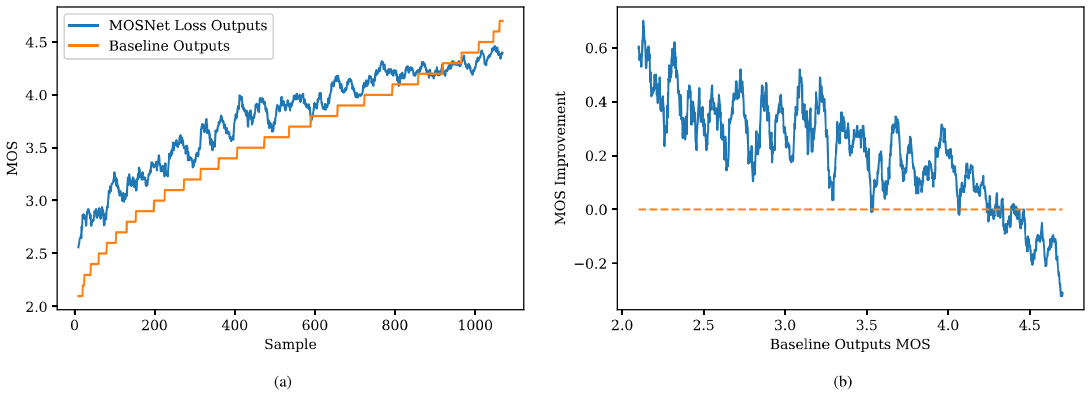

Fig. 3. Subjective evaluation of our proposed application on testset. (a) MOS ratings of outputs of the speech enhancement models and the noisy samples in ascending order with respect to MOS ratings of the baseline outputs, (b) difference between MOS ratings of our proposed model and the baseline model (blue curve − orange curve).

4.2.2. Evaluation

For evaluation, we generate a test set from the DNS challenge [18] consisting of 1080 pairs of noisy and corresponding clean 10-second speech signals with SNR levels uniformly distributed in the range $[-1, 24]$ dB. We evaluate the performance of both models using both objective and subjective metrics. For subjective evaluation, we share speech samples from the test set. We observed that while the performance of noise attenuation in both the baseline and our proposed model is competitive, the speech quality of our proposed model output is significantly better compared to the baseline output. To quantify the performance in terms of subjective metrics, we conduct ITU-T P.808 tests on AMT to acquire accurate human MOS ratings. For each record, we get MOS ratings for three signals: the noisy speech, the output of the baseline model, and the output of the MOSNet loss model. For visualization, we sort the samples in ascending order with respect to the MOS ratings of the samples of the baseline outputs. Fig. 3(a) shows the MOS ratings of the baseline outputs and the MOSNet loss model outputs. The sorted plots of our model's outputs are smoothed using a moving average with a window of size 20. For further visualization, the difference between the MOS ratings of our model and the baseline model is presented in Fig. 3(b). We observe an average improvement of 0.22 in terms of MOS. As can be seen, when the baseline model produces enhanced speech with low MOS ratings, we observe significant improvement in the perceptual quality of our proposed model's outputs.

Objective evaluation in terms of PESQ and STOI is presented in Fig. 4. Here as well, the samples are sorted in ascending order with respect to the MOS ratings of the baseline outputs. The sorted plots are smoothed using a moving average with a window of size 20. Interestingly, we observe that while PESQ and STOI are relatively correlated to the MOS ratings (the plots in

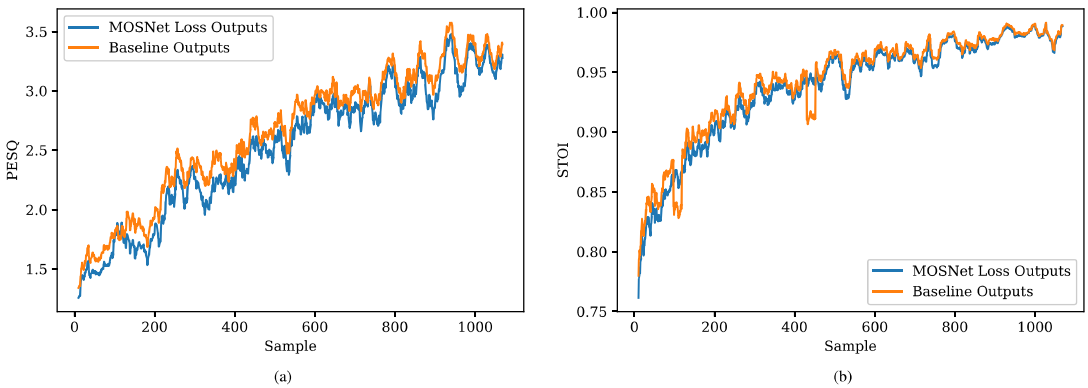

Fig. 4. Objective evaluation of our proposed application on testset in terms of (a) PESQ and (b) STOI scores.

Figs. 4 and 3(a) indicate that PESQ and STOI are relatively monotonic increasing with respect to the sorted samples), but there is a small degradation in performance compared to the baseline. Previous research suggested that traditional objective metrics are sensitive to perceptually-invariant transformations [5,6]. However, in our case, we observe significant improvement in MOS ratings despite the degradation in performance in terms of PESQ and STOI. More specifically, we observe an average degradation of 0.01 in terms of STOI and 0.13 in terms of PESQ. Furthermore, we listened to the output audio samples and observed that our proposed model significantly reduces distortions to the speech compared to the baseline model. In terms of noise attenuation, however, both models maintain similar performance level. This interesting observation suggests that PESQ and STOI metrics might be more sensitive to differences in noise attenuation performance compared to degradation in speech naturalness.

5. Conclusions

We have proposed a deep-learning-based metric to evaluate the perceptual quality of speech utterances and showed how it could be leveraged to improve existing speech enhancement models. The proposed method significantly improves the perceptual quality of the denoiser outputs, especially when the speech inputs are very challenging. The perceptual regularization effectively reduces distortions to the speech while maintaining a similar noise attenuation performance level. While we focused on speech denoising, the concept of perceptual regularization can be extended to other fields of speech processing. An interesting direction for future research is investigating the sensitivity of the proposed metric to other speech perturbations. Intuitively, it makes sense that noisy speech with a low SNR level will also have a low MOS rating, and that improved perceptual quality is more related to the level of speech distortion than to the level of noise attenuation. However, this intuition is unclear when considering perturbations such as reverberation and compression. It is interesting to understand which artifacts can be reduced using perceptual regularization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

DNS Challenge (Reference data) (github).

References

[1] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings, volume 2, IEEE, 2001, pp. 749–752.

[2] J.G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, M. Keyhl, Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—temporal alignment, Journal of the Audio Engineering Society 61 (6) (2013) 366–384.

[3] A. Hines, J. Skoglund, A.C. Kokaram, N. Harte, ViSQOL: an objective speech quality model, EURASIP Journal on Audio, Speech, and Music Processing 2015 (1) (2015) 1–18.

[4] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: 2010 IEEE international conference on acoustics, speech and signal processing, IEEE, 2010, pp. 4214–4217.

[5] A. Hines, J. Skoglund, A. Kokaram, N. Harte, Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 3697–3701.

[6] T. Manjunath, Limitations of perceptual evaluation of speech quality on VoIP systems, in: 2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, IEEE, 2009, pp. 1–6.

[7] J.M. Martin-Doñas, A.M. Gomez, J.A. Gonzalez, A.M. Peinado, A deep learning loss function based on the perceptual evaluation of the speech quality, IEEE Signal Processing Letters 25 (11) (2018) 1680–1684, doi:10.1109/LSP.2018.2871419.

[8] S.-W. Fu, C.-F. Liao, Y. Tsao, S.-D. Lin, MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement, ICML, 2019.

[9] Z. Xu, M. Strake, T. Fingscheidt, Deep noise suppression maximizing non-differentiable PESQ mediated by a non-intrusive PESQNet, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022) 1572–1585.

[10] C.K. Reddy, V. Gopal, R. Cutler, DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6493–6497.

[11] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, H.-M. Wang, MOSNet: Deep learning-based objective assessment for voice conversion, in: Proc. Interspeech 2019, 2019, pp. 1541–1545, doi:10.21437/Interspeech.2019-2003.

[12] P. Manocha, A. Finkelstein, R. Zhang, N.J. Bryan, G.J. Mysore, Z. Jin, A differentiable perceptual audio metric learned from just noticeable differences, Interspeech, 2020.

[13] P. Manocha, Z. Jin, R. Zhang, A. Finkelstein, CDPAM: Contrastive learning for perceptual audio similarity, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 196–200.

[14] J. Serrà, J. Pons, S. Pascual, SESQA: semi-supervised learning for speech quality assessment, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 381–385.

[15] P. Manocha, B. Xu, A. Kumar, NORESQA: A framework for speech quality assessment using non-matching references, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 22363–22378.

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

[17] B. Naderi, T. Hoßfeld, M. Hirth, F. Metzger, S. Möller, R.Z. Jiménez, Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach, in: 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), 2020, pp. 1–6.

[18] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E.S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, R. Aichner, ICASSP 2022 deep noise suppression challenge, ICASSP, 2022.

[19] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, T. Nakatani, Neural network-based spectrum estimation for online WPE dereverberation, in: Proc. Interspeech 2017, 2017, pp. 384–388.

[20] E. Habets, J. Benesty, S. Gannot, I. Cohen, The MVDR beamformer for speech enhancement, in: Speech Processing in Modern Communication, Springer, 2010, pp. 225–254.

[21] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer-Verlag, Berlin, Germany, 2009, pp. 37–40.

[22] A. Défossez, G. Synnaeve, Y. Adi, Real time speech enhancement in the waveform domain, in: Proc. Interspeech 2020, 2020, pp. 3291–3295.

[23] A. Défossez, N. Usunier, L. Bottou, F. Bach, Music source separation in the waveform domain, arXiv preprint arXiv:1911.13254 (2019).

[24] C. Valentini-Botinhao, Noisy speech database for training speech enhancement algorithms and TTS models, 2017.