

Impairments are Clustered in Latents of Deep Neural Network-based Speech Quality Models

Fredrik Cumlin ^{*}, Xinyu Liang [†], Victor Ungureanu [‡], Chandan K. A. Reddy [‡], Christian Schüldt [‡], Saikat Chatterjee ^{*}

^{*} School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden

[†] HCLTech, Sweden [‡] Google LLC

^{*} {fcumlin, sach}@kth.se, [†] hopeliang@icloud.com, [‡] {ungureanu, chandanka, cschuldt}@google.com

Abstract—In this article, we provide an experimental observation: Deep neural network (DNN) based speech quality assessment (SQA) models have inherent latent representations where many types of impairments are clustered. While DNN-based SQA models are not trained for impairment classification, our experiments show good impairment classification results in an appropriate SQA latent representation. We investigate the clustering of impairments using various kinds of audio degradations that include different types of noises, waveform clipping, gain transition, pitch shift, compression, reverberation, etc. To visualize the clusters we perform classification of impairments in the SQA-latent representation domain using a standard k-nearest neighbor (kNN) classifier. We also develop a new DNN-based SQA model, named DNSMOS+, to examine whether an improvement in SQA leads to an improvement in impairment classification. The classification accuracy is 94% for LibriAugmented dataset with 16 types of impairments and 54% for ESC-50 dataset with 50 types of real noises.

Index Terms—latent representations, noise classification, speech quality assessment, deep neural network.

I. INTRODUCTION

Auditory-motivated perceptual features are extensively used in audio signal processing. Examples of time-tested features are Mel-frequency cepstral coefficients (MFCCs) [1], perceptual linear prediction (PLP) [2], and Gammatone frequency cepstral coefficients (GFCCs) [3]. The use of MFCC features is a standard in speech recognition, speaker identification, and many other speech processing tasks [1].

Interesting experimental evidence exists that noise sounds, such as coughing, train noise, mouse-clicking, etc., get clustered in the MFCC feature vector domain [4]. To visualize the clusters, a baseline kNN classifier was used in the MFCC feature space, to correctly classify noise types 66.7% in ESC-10 and 32.2% in ESC-50. Note that the signal processing steps to generate MFCC feature vectors are auditory-motivated, and not explicitly designed for noise classification. Therefore, the experimental evidence suggests that noise classification happens in an auditory-motivated feature domain (the MFCC domain).

Speech quality assessment (SQA) is the task of estimating speech quality on a scale based on human perception. Speech quality is affected by different impairments, such as varying types of noises, reverberation, codec artifacts, etc. Currently, the best-performing SQA models are designed using deep neural networks (DNNs) [5], [8], [9], [11], [13], [16].

In this article, we show a phenomenon: DNN-based SQA models produce internal (latent) feature vectors where many types of impairments are clustered. Note that SQA models are not engineered for clustering impairments. The training and learning algorithms for the DNN-based SQA methods do not use any data and/or optimization

problems related to impairment clustering. The clustering of impairments turns out to be an inherent by-product. To the best of the authors' knowledge, our work is the first experimental evidence of the phenomenon. Our contributions are as follows:

- 1) To visualize the clustering of impairments in a suitable latent feature space, we perform a classification task to identify the types of impairments. For this, we employ a standard k-Nearest Neighbors (kNN) method, similar to the approach used by [4]. Our experiments demonstrate that the impairment classification accuracy is significantly higher in the latent domain of DNN-based SQA models compared to the traditional MFCC domain. For instance, in the ESC-50 dataset, which contains 50 distinct noise types, we achieve a classification accuracy of 27% in the latent domain of a DNN-based SQA model, compared to just 18% in the MFCC domain.
- 2) We hypothesize that an improvement in SQA performance correlates with an enhancement in impairment classification accuracy. To test this hypothesis, we first evaluate impairment classification accuracy in the latent domain of an existing DNN-based SQA model, DNSMOS [6]. We then apply an improved version, DNSMOS Pro [7], and examine its impact on impairment classification. To adapt DNSMOS Pro for a regression task (instead of the probabilistic approach proposed in the original paper), we introduce slight modifications and refer to the resulting model as DNSMOS+. Using the LibriAugmented dataset, we observe that DNSMOS achieves an SQA performance of 0.89, as measured by the Pearson correlation coefficient (PCC), with an impairment classification accuracy of 86%. The enhanced DNSMOS+ model improves SQA performance to 0.93 (PCC), with a corresponding increase in impairment classification accuracy to 94%, thereby supporting our hypothesis.

A. DNN-based Speech Quality Assessment Models

In this subsection, we provide a brief literature review on SQA models, primarily on non-intrusive DNN-based SQA models.

For a non-intrusive SQA, the goal is to predict the speech quality given only a degraded speech signal (i.e., in the absence of a reference signal). The SQA models also can be intrusive, for example, PESQ [17], ViSQOL [18].

Over the past years, DNN-based non-intrusive SQA models have become dominant. Early works include AutoMOS [12] and QualityNet [10], where the first is trained using mean-opinion-score (MOS) [20] and the latter is trained using PESQ [17] labels. Later works have been trained end-to-end with MOS labels, such as DNSMOS, or utilizing individual scores for increased performance [5], [8], [11]. Further, there are unsupervised feature extraction-based SQA models [13], [15], [21], [22].

The computations handling was enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation.

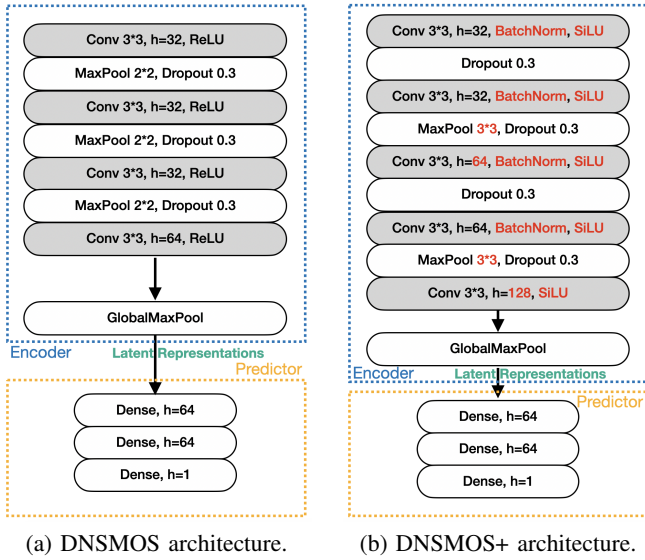


Fig. 1. Architectures of DNSMOS and DNSMOS+.

Some methods have explored measuring the quality from more than one perspective, such as DNSMOSp835 [23] and NISQA [9]. DNSMOSp835 is trained on scores collected under ITU-T Rec. P.835 [19], which aims to emulate the subjective evaluation for the amount of noise and the impairment of the speech together with the overall speech quality. NISQA hypothesizes that overall speech quality can be described by four different quality 'dimensions': noisiness, coloration, discontinuity, and loudness. However, none of the mentioned SQA models provides the specific *cause* of poor speech quality, instead, only scalar estimates of the quality are provided.

In this article, we use the DNSMOS models [6], [23] to study how impairments are represented in latent space. The choice of DNSMOS is due to its low complexity. Then we also study how the impairments are represented in the slightly improved model DNSMOS Pro [7].

II. METHOD

We start with the basic DNSMOS [6] model, which is a popular low-complexity non-intrusive SQA method. It uses the Mel spectrogram as input and outputs a scalar value: the speech quality in terms of MOS. The DNSMOS architecture consists of two parts, an encoder, and a predictor, as shown in Figure 1 (a). The encoder consists mainly of convolutional and pooling layers and produces latent representations of the input spectrogram. The predictor consists of dense layers and maps the latent representation to a scalar MOS value. The training of DNSMOS is an end-to-end supervised learning mechanism where labeled data is used. The labeled data is comprised of a pairwise input spectrogram and the corresponding MOS output as a label.

A. Clustering of impairments in latent representation

A DNSMOS predictor provides SQA in terms of MOS using the latent representation. If a speech signal has no impairments then the MOS output is expected to be good. On the other hand, the presence of impairments is expected to provide a poor MOS value. That means the latent representation has some information about the impairments. Then natural questions arise:

- 1) How is the information about the impairments reflected in the latent representation space?

- 2) Can we visualize the effect of impairments in the latent representation space?
- 3) Do different kinds of impairments form clusters in the latent space? If so, are the clusters distinct or highly overlapping?

The dimension of the latent representation vector is much smaller than the input spectrogram dimension. In our case, we use latent dimension as either 64 or 128. Therefore, visualization of the latent representation of a speech signal under an impairment or multiple types of impairments is not trivial.

An indirect approach to visualize the latent representation and forming of clusters is via a classification study: we create a labeled dataset of speech using different types of impairments and measure the accuracy of a kNN classifier.

B. Dataset construction

For a more controlled experimental setup, we also generate our pseudo-labeled dataset and train a non-intrusive SQA model end-to-end from speech clip features to speech quality. The dataset is given by $\mathcal{D} = (\mathbf{x}, y, z)_{\mathbf{x} \in \mathcal{X}}$, where \mathbf{x} is an impaired signal, y is speech quality label from an intrusive quality measure, and z is the impairment class applied to a clean speech signal to produce \mathbf{x} .

We use the LibriSpeech [24] dataset as the source of clean speech data, and introduce different types of impairments to the speech clips. This clean speech dataset consists of 100 hours of speech, a total of 28 539 speech clips, with an average duration of 13 s and sample rate 16 kHz. We crop or repetitively pad all clean speech clips to 10 s.

The impairments are added through augmentations supported by the Audiomentations library¹. We corrupt the clean signals with 9 different single impairments and 6 double impairments to make our simulated dataset closer to real-world scenarios. The chosen ratio of impairments as well as the parameters, are listed in Table I. For the additive background noise, we randomly sample it from Demand² and FreeSound³, which in total consists of 23 different noise data sources. We call the constructed dataset **LibriAugmented**, consisting of 57 078 speech clips and 16 impairment classes (considering the identity mapping as a class). The subset of LibriAugmented in which speech clips are corrupted by at most one impairment is noted as LibriAugmented_{one}. The complete LibriAugmented contains each clean speech clip from LibriSpeech exactly twice.

We label the dataset using PESQ [17] and ViSQOL v3 [18], both of which are designed to simulate MOS measures. We split the data into three sets: a training dataset, a validation dataset, and a testing dataset. We explicitly make the split between one-impairment and two-impairment subsets separately, so that each half contains 22 539, 3 000, and 3 000 speech clips in train/val/test respectively. The clean speech utterances are disjoint in the train/val/test split, meaning that for example if we have a speech clip in the one-impairment train set, then the same clean speech (but distorted with two impairments) is also in the two-impairment train set.

C. DNSMOS+ design

We use an improved DNSMOS model, namely DNSMOS Pro [7], to study how improvements in quality assessment relate to the clustering of impairments. We change the model slightly from the probabilistic definition given in [7], and we call this model DNSMOS+. Changes of DNSMOS+ architecture from the DNSMOS

¹<https://github.com/iver56/audiomentations>

²<https://zenodo.org/records/1227121>

³<https://freesound.org>

TABLE I
THE LIBRIAUGMENTED DATASET.

| Ratio | Impairment (Parameters) |
|-------|--|
| 0.1 | Identity |
| 0.1 | AddBackgroundNoise (snr=-10 ~ 15 dB) |
| 0.1 | ClippingImpairment (percentile=10 ~ 40%) |
| 0.1 | GainTransition (gain=-60 ~ 20 dB) |
| 0.1 | LowPassFilter (cutoff_freq=0.5 ~ 1 kHz) |
| 0.1 | Mp3Compression (bit_rate=8 ~ 14) |
| 0.1 | PitchShift (semitones=-4 ~ 4 kHz) |
| 0.1 | RoomSimulator (rt60=0.8 ~ 1.5 s) |
| 0.1 | TimeMask (band_part=0.2 ~ 0.5) |
| 0.1 | TimeStretch (rate=0.5 ~ 2) |
| 0.167 | AddBackgroundNoise + RoomSimulator |
| 0.167 | AddBackgroundNoise + LowPassFilter |
| 0.167 | AddBackgroundNoise + TimeStretch |
| 0.167 | RoomSimulator + Mp3Compression |
| 0.167 | PitchShift + LowPassFilter |
| 0.167 | GainTransition + TimeMask |

architecture include (1) STFT preprocessing instead of Mel spectrogram as input, (2) an extra convolutional layer, (3) SiLU (Swish) activation function instead of ReLU, (4) more sparse max-pooling layers with larger kernels, and (5) the addition of batch normalization layers. A comparison of the architectures is illustrated in Figure 1.

III. EXPERIMENTS

A. Datasets

In our experiments, we use two different datasets: the LibriAugmented dataset as described in II-B, and the Environmental Sound Classification 50 (ESC-50) dataset [4]. LibriAugmented is used as a training dataset for the non-intrusive models, and to study the clustering of different impairments. ESC-50 consists of environmental noise files and is used to study the clustering of noises in the latent space of a non-intrusive SQA model.

The ESC-50 dataset consists of 50 different noise categories with 40 samples in each category, resulting in a total of 2 000 noise clips. All signals are 5 s long and sampled at 16 kHz. Examples include noise types are mouse-clicking, cow, thunderstorm, laughing, vacuum cleaner, baby crying, coughing, and helicopter. Human accuracy on this dataset is $\sim 81.3\%$ [4].

B. Feature extraction

The feature extraction of the speech clips, before processing by the DNN, is similar to DNSMOS [6]. The inputs are log-magnitude spectrograms with a window duration of 20 ms and a hop duration of 10 ms, using a Hann window, and at a sample rate of 16 kHz. Furthermore, the values in the spectrogram are clipped to the interval $[-7, 7]$. This is to avoid having large values making training unstable.

C. Experimental setup

We design four training experiments: (1) DNSMOS (architecture) on LibriAugmented dataset using ViSQOL labels; (2) DNSMOS+ on LibriAugmented_{one} using ViSQOL labels; (3) DNSMOS+ on LibriAugmented using PESQ labels; (4) DNSMOS+ on LibriAugmented using ViSQOL labels.

The experiments were designed from an observational point of view, where the latent clustering is studied in these configurations⁴.

For all experiments, we train end-to-end from speech clip features to quality labels. We use the mean-square-error (MSE) as a loss function. An Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [14]

⁴Code can be found at https://github.com/fcumlins/sqa_latent_classification.

was used, together with a learning rate of 10^{-4} and batch size of 32, as per DNSMOS [6]. We train for 500 epochs and extract the model for analysis once it has been fully trained. One epoch means one iteration over the whole training dataset. Training is done on one Nvidia A100 40GB GPU card, and training one model takes ~ 30 hours to complete. Testing is done on the LibriAugmented test data.

1) *Performance on predicting quality label*: As performance measures we report the mean squared error (MSE), Pearson Correlation Coefficient (PCC), and Spearman Rank Correlation Coefficient (SRCC). These measurements are used to compare the predictions with the ViSQOL labels on the test data. If the model is trained on PESQ labels, we use the PESQ labels on the test data instead.

2) *Performance on predicting impairments*: To measure the latent clustering of the impairments we do as explained in the method Section II. Given the LibriAugmented test dataset, we have impaired speech clips and impairment class label pairs. Some speech clips have been distorted by two impairments, and we consider each such pair as its own class. This means each speech clip has exactly one impairment label.

We further split this LibriAugmented test set for training and testing a kNN model, and this split is done in a stratified way so that the proportion of different classes in each split is approximately the same. We use 70% of data in training and 30% in test. A kNN, using $k = 15$, is trained on the latent representations of the training subset. We measure the accuracy of predicting the augmentation on the test subset.

3) *Performance on predicting noise class*: To measure the classification performance of different noises we use the ESC-50 dataset. We partition this dataset into two sets, train and test, in a similar stratified partitioning procedure as described in the previous paragraph. A kNN, using $k = 15$, is also trained similarly, and we measure the accuracy of predicting the noise class on the test subset.

4) *Baseline and models from the literature*: For the classification task of predicting the impairment class in the LibriAugmented dataset and the noise class in the ESC-50 dataset, we use three additional methods for comparison.

As a baseline, we consider random projection [25]. This is done by initializing a matrix A in $\mathbb{R}^{d \times 128}$, where d is the number of samples of an audio clip. The initialization is done so the columns in the matrix are realizations of independent and identically distributed normal vectors with an expected l_2 length of 1. The mapping of an audio clip x to a 128-dimensional vector is given by $x \mapsto Ax$. Subsequently, classification analysis is done using the image of a speech clip under the random projection as input to the kNN model.

We also include classification results when using MFCC as an input feature. This is done by computing the MFCCs of the speech clips using 12 bins and using the image as input to the explained kNN procedure.

We furthermore study pre-trained models in the literature, namely DNSMOS [6] and DNSMOSp835 [23]⁵. The models were extracted as-is; no additional training of the model weights was done. The latent representations after the GlobalMaxPool layer were used as input to train a kNN as explained.

D. Results

The results of speech quality prediction performance and classification performance are shown in Table II. Note that the main interest is the observation of the clustering of impairments, not speech quality

⁵The models can be found in <https://github.com/microsoft/DNS-Challenge/tree/master/DNSMOS>.

TABLE II

PERFORMANCE RESULTS ON LIBRIAUGMENTED AND ESC-50. ALL SELF-IMPLEMENTED ALGORITHMS ARE RUN 10 TIMES WITH DIFFERENT SEEDS, AND THE MEAN AND STANDARD DEVIATION ARE REPORTED. WHEN TRAINED TOWARDS PESQ LABELS, WE ALSO EVALUATE ON PESQ LABELS ON THE LIBRIAUGMENTED DATASET.

| Model | Labels | Training data | LibriAugmented | | | | ESC-50 | |
|-------------------|--------|-------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | | | MSE | PCC | SRCC | Top-1 Acc | Top-1 Acc | Top-3 Acc |
| Random projection | - | - | - | - | - | 0.08±0.01 | 0.03±0.01 | 0.07±0.01 |
| MFCC | - | - | - | - | - | 0.30 | 0.18 | 0.36 |
| DNSMOS | MOS | DNS data [6] | 1.13 | 0.41 | 0.43 | 0.59 | 0.27 | 0.52 |
| DNSMOSp835 | MOS | DNS data [23] | 0.80 | 0.59 | 0.64 | 0.74 | 0.27 | 0.53 |
| DNSMOS | ViSQOL | LibriAugmented | 0.28±0.01 | 0.89±0.01 | 0.90±0.01 | 0.86±0.01 | 0.42±0.02 | 0.67±0.01 |
| DNSMOS+ | ViSQOL | LibriAugmented _{one} | 0.21±0.04 | 0.90±0.01 | 0.89±0.02 | 0.93±0.01 | 0.51±0.02 | 0.78±0.02 |
| DNSMOS+ | PESQ | LibriAugmented | 0.33±0.15 | 0.87±0.05 | 0.87±0.01 | 0.93±0.01 | 0.49±0.02 | 0.76±0.01 |
| DNSMOS+ | ViSQOL | LibriAugmented | 0.19±0.03 | 0.93±0.01 | 0.94±0.01 | 0.94±0.01 | 0.54±0.01 | 0.80±0.01 |

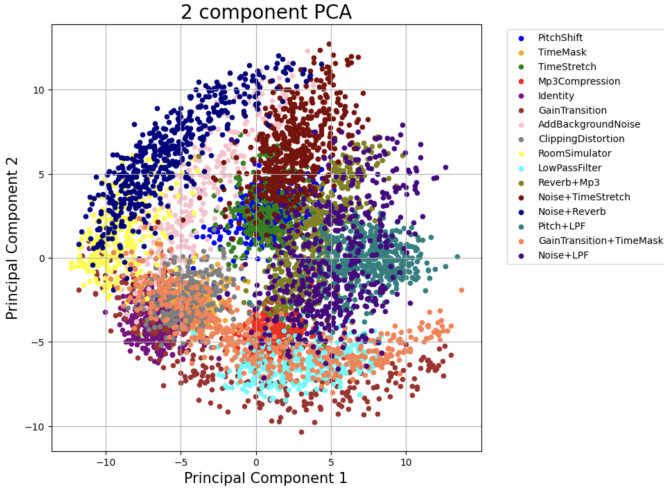


Fig. 2. PCA visualization, using output dimension of 2, of latent features of the LibriAugmented datasets using DNSMOS+ (ViSQOL labels).

prediction performance per se. In Fig. 2, a 2-component PCA of the latent representation of LibriAugmented dataset in the DNSMOS+ model is shown.

We notice that models trained on LibriAugmented and models from literature outperform random projection in classifying the impairment and noise class respectively. This gives evidence of an inductive bias (“tendency”) of the DNNs to cluster different impairments.

Furthermore, we can see a clear difference in the noise classification accuracy when comparing the DNSMOS models from the literature to the LibriAugmented trained ones. One reason for this can be about the data: The DNS dataset consists of 600 noisy speech clips processed by > 200 noise suppressors and 40 noise suppressors for DNSMOS and DNSMOSp835 respectively. Our training dataset consists of 22 539 unique speech clips processed with one or two impairments, which might increase the clustering of noises. Another plausible explanation is that the DNS data used has been processed by noise suppressors, and so could suppress the noises to a degree where clustering of noises is exhibited to a smaller extent. The reason herein is an interesting topic for future study.

Considering the trained models, it shows that DNSMOS+ is a more suitable architecture than DNSMOS. One reason could be the increased dimension of the latent space (64-dimensional vector for DNSMOS to 128-dimensional vector for DNSMOS+). This poten-

tially increases the expressivity for DNSMOS+. Suitably applying normalization and other changes makes DNSMOS+ have higher quality prediction measures, and also higher classification accuracy.

We also notice that both ViSQOL and PESQ labels demonstrate their suitability for being used as labels in the SQA task and the classification task (the last two rows in Table II). However, ViSQOL labels seem more suitable compared to PESQ labels, as the average performance is higher with a lower standard deviation across runs when using ViSQOL labels compared to PESQ labels. The reason could be that ViSQOL(v3) is a better MOS emulator compared to PESQ for the considered impairments and datasets, which is consistent with the authors’ subjective opinions when listening to the created dataset; ViSQOL is more aligned with our judgment compared to PESQ.

In Fig. 2, a 2-component PCA visualization of the latent representations from the LibriAugmented dataset using the DNSMOS+ model is presented. The labels correspond to the distortions applied to the signals before processing. The plot reveals that speech signals with similar distortions tend to cluster together. Importantly, this clustering is not driven by the quality labels, as the quality distributions are roughly equal across different distortion classes. Therefore, the model’s ability to cluster these distortions appears to be an emergent behavior, rather than a consequence of any single distortion consistently producing higher or lower quality signals.

IV. CONCLUSION

In this paper, we provided an experimental observation that impairments are clustered in the latent space of DNN-based non-intrusive speech quality models. These models are trained end-to-end to predict only speech quality scores. Using DNSMOS from the literature, we showed 30% top-1 noise accuracy and 60–75% top-1 impairment accuracy. Using a regression-suitable DNSMOS Pro architecture and training on a more diverse dataset, we achieved 54% top-1 noise accuracy and 94% top-1 impairment accuracy. Important to note is that *the models were not trained for impairment classification* - they were only trained to predict quality. The results, together with the low accuracy when using random linear projection, support the hypothesis that DNN-based SQA models have inherent latent representations where many types of impairments are clustered.

REFERENCES

- [1] Z. K. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and its Applications: A Review,” *IEEE Access*, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.
- [2] F. Hönl, G. Stemmer, C. Hacker, and F. Brugnara, “Revising Perceptual Linear Prediction (PLP),” in *Proc. Interspeech*, 2005.

- [3] B. Ayoub, J. Kharroubi, and A. Zarghili, "Gammatone frequency cepstral coefficients for speaker identification over VoIP networks," in *Proc. 2016 Int. Conf. Information Technology for Organizations Development (IT4OD)*, 2016, pp. 1-5, doi: 10.1109/IT4OD.2016.7479293.
- [4] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. 23rd Annual ACM Conf. Multimedia*, Brisbane, Australia, Oct. 2015, pp. 1015-1018, doi: 10.1145/2733373.2806390.
- [5] F. Cumlin, C. Schüldt, and S. Chatterjee, "Latent-based Neural Net for Non-intrusive Speech Quality Assessment," in *Proc. 2023 33rd European Signal Processing Conference (EUSIPCO)*, Sept. 2023, pp. 36-40.
- [6] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, ON, Canada, June 6-11, 2021, doi: 10.1109/ICASSP39728.2021.9414878.
- [7] F. Cumlin, X. Liang, V. Ungureanu, C. Reddy, C. Schüldt, S. Chatterjee, "DNSMOS Pro: A Reduced Size DNN for Probabilistic MOS of Speech," in *Proc. Interspeech 2024, 25th Annu. Conf. Int. Speech Commun. Assoc.*, Kos Island, Greece, Sept. 1-5, 2024.
- [8] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, doi: 10.1109/ICASSP43922.2022.9747222.
- [9] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech 2021*, Aug. 2021, doi: 10.21437/Interspeech.2021-299.
- [10] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM," in *Proc. Interspeech 2018, 19th Annu. Conf. Int. Speech Commun. Assoc.*, Hyderabad, India, Sept. 2-6, 2018, doi: 10.21437/Interspeech.2018-1802.
- [11] X. Liang, F. Cumlin, C. Schüldt, and S. Chatterjee, "DeePMOS: Deep Posterior Mean-Opinion-Score of Speech," in *Proc. Interspeech 2023*, pp. 526-530.
- [12] B. Patton, Y. Agiomyriannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a Non-Intrusive Assessor of Naturalness-of-Speech," in *Proc. NIPS 2016 End-to-End Learning for Speech and Audio Processing Workshop*, 2016.
- [13] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SARULAB System for VoiceMOS Challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.
- [14] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [15] Z. Qi, X. Hu, W. Zhou, S. Li, H. Wu, J. Lu, and X. Xu, "LE-SSL-MOS: Self-Supervised Learning MOS Prediction with Listener Enhancement," in *Proc. 2023 IEEE Int. Conf. Multimedia and Expo (ICME)*, 2023.
- [16] A. Stan, "The Zevomos Entry to VoiceMOS Challenge 2022," *arXiv preprint arXiv:2206.07448*, 2022.
- [17] 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 *Proc. ICASSP 2001*, vol. 2, 2001, doi: 10.1109/ICASSP.2001.941023.
- [18] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," *arXiv*, 2020, doi: 10.48550/ARXIV.2004.09584.
- [19] ITU-T Recommendation P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," International Telecommunication Union, Geneva, 2003.
- [20] ITU-T Recommendation P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," International Telecommunication Union, Geneva, 2018.
- [21] O. Platek and O. Dusek, "MooseNet: A Trainable Metric for Synthesized Speech with a PLDA Module," in *Proc. 12th Speech Synthesis Workshop (SSW)*, 2023.
- [22] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," in *Proc. ICASSP 2022*, 2022, pp. 8442-8446, doi: 10.1109/ICASSP43922.2022.9746395.
- [23] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP 2022*, 2022, pp. 886-890.
- [24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP 2015*, 2015, doi: 10.1109/ICASSP.2015.7178964.
- [25] I. Kononenko and M. Kukar, "Data preprocessing," in *Machine Learning and Data Mining*, I. Kononenko and M. Kukar, Eds. Woodhead Publishing, 2007, pp. 181-211, doi: 10.1533/9780857099440.181.